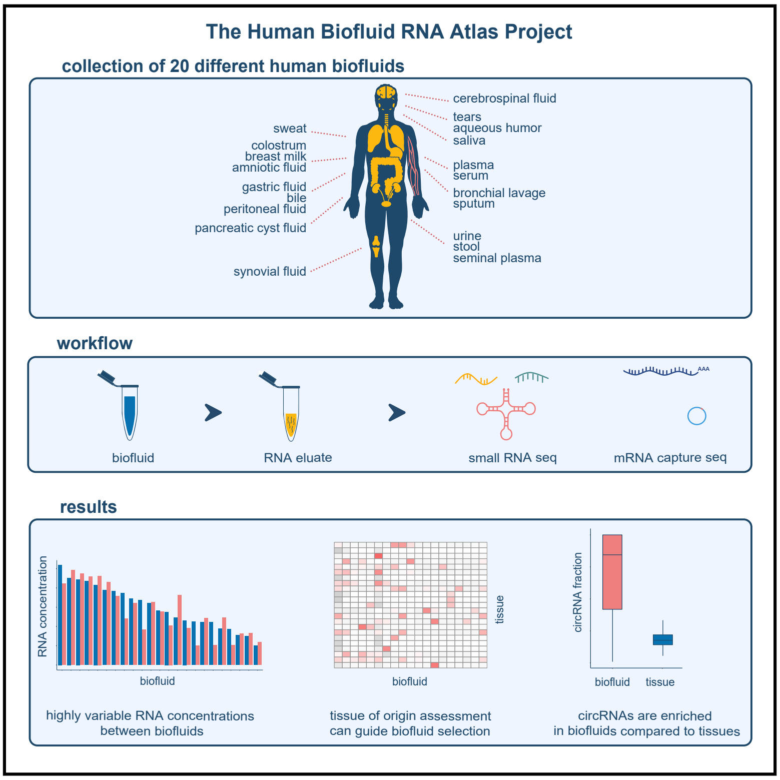
Cell Reports

Resource

Charting Extracellular Transcriptomes in The Human Biofluid RNA Atlas

Graphical Abstract



Highlights

- Messenger, circular, and small RNA transcriptomes of 20 biofluids are studied
- Synthetic spike-in controls allow direct comparison between biofluids
- Tissues of origin are assessed per biofluid
- The results can guide the selection of biofluids in biomarker research

Authors

Eva Hulstaert, Annelien Morlion, Francisco Avila Cobos, ..., Gary P. Schroth, Jo Vandesompele, Pieter Mestdagh

Correspondence

pieter.mestdagh@ugent.be

In Brief

Hulstaert et al. present an atlas of messenger, circular, and small RNA transcriptomes of 20 biofluids. Spike-in RNA controls allow direct comparison of the RNA content between the biofluids. Tissues of origin are assessed per biofluid and can guide the selection of the most appropriate biofluid for future biomarker studies.

Hulstaert et al., 2020, Cell Reports *33*, 108552
December 29, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.celrep.2020.108552

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

Charting Extracellular Transcriptomes in The Human Biofluid RNA Atlas

Eva Hulstaert,[1,2,3] Annelien Morlion,[1,2] Francisco Avila Cobos,[1,2] Kimberly Verniers,[1,2] Justine Nuytens,[1,2] Eveline Vanden Eynde,[1,2] Nurten Yigit,[1,2] Jasper Anckaert,[1,2] Anja Geerts,[4] Pieter Hindryckx,[4] Peggy Jacques,[5,6] Guy Brusselle,[7] Ken R. Bracke,[7] Tania Maes,[7] Thomas Malfait,[7] Thierry Derveaux,[8] Virginie Ninclaus,[8] Caroline Van Cauwenbergh,[8] Kristien Roelens,[9] Ellen Roets,[9] Dimitri Hemelsoet,[10] Kelly Tilleman,[11] Lieve Brochez,[2,3] Scott Kuersten,[12] Lukas M. Simon,[13] Sebastian Karg,[14] Alexandra Kautzky-Willers,[15] Michael Leutner,[15] Christa Nöhammer,[16] Ondrej Slaby,[17,18,19] Roméo Willinge Prins,[20] Jan Koster,[20] Steve Lefever,[1,2] Gary P. Schroth,[12] Jo Vandesompele,[1,2,21] and Pieter Mestdagh[1,2,21,22,*]

[1]Center for Medical Genetics, Department of Biomolecular Medicine, Ghent University, C. Heymanslaan 10, 9000 Ghent, Belgium
[2]OncoRNALab, Cancer Research Institute Ghent (CRIG), Ghent University, C. Heymanslaan 10, 9000 Ghent, Belgium
[3]Department of Dermatology, Ghent University Hospital, C. Heymanslaan 10, 9000 Ghent, Belgium
[4]Department of Gastroenterology, Ghent University Hospital, C. Heymanslaan 10, 9000 Ghent, Belgium
[5]Department of Rheumatology, Ghent University Hospital, C. Heymanslaan 10, 9000 Ghent, Belgium
[6]VIB Inflammation Research Center, Ghent University, Technologiepark 71, 9052 Ghent, Belgium
[7]Department of Respiratory Medicine, Ghent University Hospital, C. Heymanslaan 10, 9000 Ghent, Belgium
[8]Department of Ophthalmology, Ghent University Hospital, C. Heymanslaan 10, 9000 Ghent, Belgium
[9]Department of Obstetrics, Women's Clinic, Ghent University Hospital, C. Heymanslaan 10, 9000 Ghent, Belgium
[10]Department of Neurology, Ghent University Hospital, C. Heymanslaan 10, 9000 Ghent, Belgium
[11]Department of Reproductive Medicine, Ghent University Hospital, C. Heymanslaan 10, 9000 Ghent, Belgium
[12]Illumina, San Diego, CA, USA
[13]University of Texas Health Science Center, School of Biomedical Informatics, Center for Precision Health, Houston, TX, USA
[14]Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany
[15]Department of Internal Medicine III, Clinical Division of Endocrinology and Metabolism, Unit of Gender Medicine, Medical University of Vienna, Vienna, Austria
[16]Austrian Institute of Technology, Center for Health and Bioresources, Molecular Diagnostics, Vienna, Austria
[17]Masaryk Memorial Cancer Institute, Department of Comprehensive Cancer Care, Brno, Czech Republic
[18]Department of Pathology, University Hospital Brno, Brno, Czech Republic
[19]Central European Institute of Technology, Masaryk University, Brno, Czech Republic
[20]Department of Oncogenomics, Amsterdam University Medical Centers (AUMC), University of Amsterdam, Amsterdam, the Netherlands
[21]These authors contributed equally
[22]Lead Contact
*Correspondence: pieter.mestdagh@ugent.be

https://doi.org/10.1016/j.celrep.2020.108552

SUMMARY

Extracellular RNAs present in biofluids have emerged as potential biomarkers for disease. Where most studies focus on blood-derived fluids, other biofluids may be more informative. We present an atlas of messenger, circular, and small RNA transcriptomes of a comprehensive collection of 20 human biofluids. By means of synthetic spike-in controls, we compare RNA content across biofluids, revealing a 10,000-fold difference in concentration. The circular RNA fraction is increased in most biofluids compared to tissues. Each biofluid transcriptome is enriched for RNA molecules derived from specific tissues and cell types. Our atlas enables an informed selection of the most relevant biofluid to monitor particular diseases. To verify the biomarker potential in these biofluids, four validation cohorts representing a broad spectrum of diseases were profiled, revealing numerous differential RNAs between case and control subjects. Spike-normalized data are publicly available in the R2 web portal for further exploration.

INTRODUCTION

Extracellular RNAs (exRNAs) in blood and other biofluids are emerging as potential biomarkers for a wide range of diseases (Freedman et al., 2016; Godoy et al., 2018; Max et al., 2018; Weiland et al., 2012; Yuan et al., 2016). These so-called liquid biopsies may offer a non-invasive alternative to tissue biopsies for both diagnosis and treatment response monitoring.

Previous studies have extensively profiled the small RNA content of several biofluids and identified large differences in

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

the small RNA content among different biofluids (El-Mogy et al., 2018; Fehlmann et al., 2016; Ferrero et al., 2017; Freedman et al., 2016; Godoy et al., 2018; Max et al., 2018; Murillo et al., 2019; Srinivasan et al., 2019; Umu et al., 2018; Weiland et al., 2012; Yeri et al., 2017; Yuan et al., 2016). These efforts were gathered by the NIH Extracellular RNA Communication Consortium in the exRNA Atlas Resource (https://exrna-atlas.org) (Murillo et al., 2019). Besides microRNAs (miRNAs), the most studied small RNA biotype in biofluids, other small RNAs, such as piwi-interacting RNAs (piRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), ribosomal RNAs (rRNAs), transfer RNA (tRNA) fragments, and Y-RNAs have also been identified (El-Mogy et al., 2018; Ferrero et al., 2017; Godoy et al., 2018; Srinivasan et al., 2019; Weber, 2017; Yeri et al., 2017). Weber (2017) was the first to compare the miRNA content in 12 different human biofluids (pooled samples of plasma, saliva, tears, urine, amniotic fluid, colostrum, breast milk, bronchial lavage fluid, cerebrospinal fluid [CSF], peritoneal fluid, pleural fluid, and seminal plasma) using quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) of selected miRNAs. Large variations in RNA concentration were observed among the different biofluids, with the highest small RNA concentrations measured in breast milk and seminal fluid. Since the advent of small RNA sequencing, other small RNA biotypes were characterized in various biofluids, such as plasma, serum, stool, urine, amniotic fluid, bronchial lavage fluid, bile, CSF, saliva, seminal plasma, and ovarian follicle fluid (El-Mogy et al., 2018; Ferrero et al., 2017; Godoy et al., 2018; Srinivasan et al., 2019).The distribution of small RNA biotypes clearly varies across these biofluids, with a high abundance of piRNAs and tRNAs reported in urine and a high abundance of Y-RNAs in plasma (El-Mogy et al., 2018; Ferrero et al., 2017; Yeri et al., 2017). Also non-human RNA sequences, mapping to bacterial genomes, were reported in plasma, urine, and saliva (Yeri et al., 2017).

A systematic RNA-sequencing analysis of biofluids to explore the messenger RNAs (mRNAs) and circular RNA (circRNA) transcriptome is challenging due to low RNA concentration and RNA fragmentation in biofluids. As such, most studies have explored the abundance of individual mRNAs in one specific biofluid by qRT-PCR (Herring et al., 2018; Maker et al., 2019; Marzioni et al., 2015; Oreo et al., 2014; Tian et al., 2016; Welch et al., 2016). circRNAs have been reported in saliva (Bahn et al., 2015), semen (Liu et al., 2019), blood (Li et al., 2018a), and urine (Kölling et al., 2019; Vo et al., 2019). Recently, the mRNA content of plasma and serum has been investigated using dedicated sequencing approaches such as phospho-RNA-seq, small input liquid volume extracellular RNA sequencing (SILVER-seq), and the SMARTer Stranded Total RNA-Seq Kit method (Everaert et al., 2019; Giraldez et al., 2019; Metzenmacher et al., 2020; Zhou et al., 2019). Studies comparing the small RNA, mRNA, and circRNA content in a wide range of human biofluids are currently lacking and are essential to explore the biomarker potential of exRNAs.

The goal of the Human Biofluid RNA Atlas is to define the extracellular transcriptome across a wide range of human biofluids (amniotic fluid, aqueous humor, ascites, bile, bronchial lavage fluid, breast milk, CSF, colostrum, gastric fluid, pancreatic cyst fluid, plasma, saliva, seminal fluid, serum, sputum, stool, synovial fluid, sweat, tear fluid, and urine) and to assess biomarker potential in selected case-control cohorts. We used small RNA sequencing to quantify different small RNA species and present a dedicated mRNA capture sequencing workflow to simultaneously quantify mRNAs and circRNAs.

In the first phase of our study, small RNA sequencing and mRNA capture sequencing were performed in a discovery cohort of 20 different biofluids (Figure 1). The goal of this phase was to assess the technical feasibility of the methodology and to generate a comprehensive set of mRNAs, circRNAs, and small RNAs in which the contributing tissues and cell types per biofluid were assessed.

In the second phase of our study, we aimed to investigate the biological relevance of exRNAs in various biofluids. Therefore, mRNA capture sequencing was applied to four different case/control cohorts, each consisting of 16–24 samples (Figure 1). These samples included sputum samples from 8 patients with chronic obstructive pulmonary disease (COPD) versus 8 controls, urine samples from 12 bladder cancer patients versus 12 controls, CSF samples from 12 glioblastoma patients versus 12 hydrocephalus patients and saliva samples from 12 diabetes mellitus patients versus 12 controls.

The resulting catalog of extracellular transcriptomes of 185 human samples can guide researchers in the biomarker field to investigate other biofluids besides the well-studied blood-derived ones and is a first step to more dedicated mRNA and circRNA profiling of biofluids in larger cohorts.

RESULTS

RNA Spike-In Controls Enable Process Control of the RNA-Sequencing Workflow

Synthetic spike-in RNA sequences are crucial to control the process from RNA isolation to RNA sequencing, especially when working with challenging and low input material. We applied 4 different mixes of synthetic RNA spike-in controls (in total, 189 RNAs) as workflow processing and normalization controls that enable direct comparison of the RNA profiles across the different biofluids. Sequin and small RNA extraction control (RC) spikes were added before RNA isolation, whereas External RNA Controls Consortium (ERCC) spikes and small RNA library preparation (LP) spikes were added to the RNA eluate before genomic DNA (gDNA) removal (Figure 1). Of note, every spike mix consists of multiple RNA molecules of different lengths over a wide concentration range. Detailed information is provided in Methods S1. Besides normalization, the spike-in controls enabled quality control of the RNA extraction and library preparation steps in the workflow and relative quantification of the RNA yield and concentration across the different biofluids.

First, the correlation between the expected and the observed relative quantities for all four spike mixes can be used to assess quantitative linearity. In the discovery cohort, the expected and the observed relative quantities for all four spike mixes were well correlated (Pearson correlation coefficients range from 0.50 to 1.00 for Sequin spikes, 0.92 to 1.00 for ERCC spikes, 0.44 to 0.98 for RC spikes, and 0.40 to 0.96 for LP spikes). In

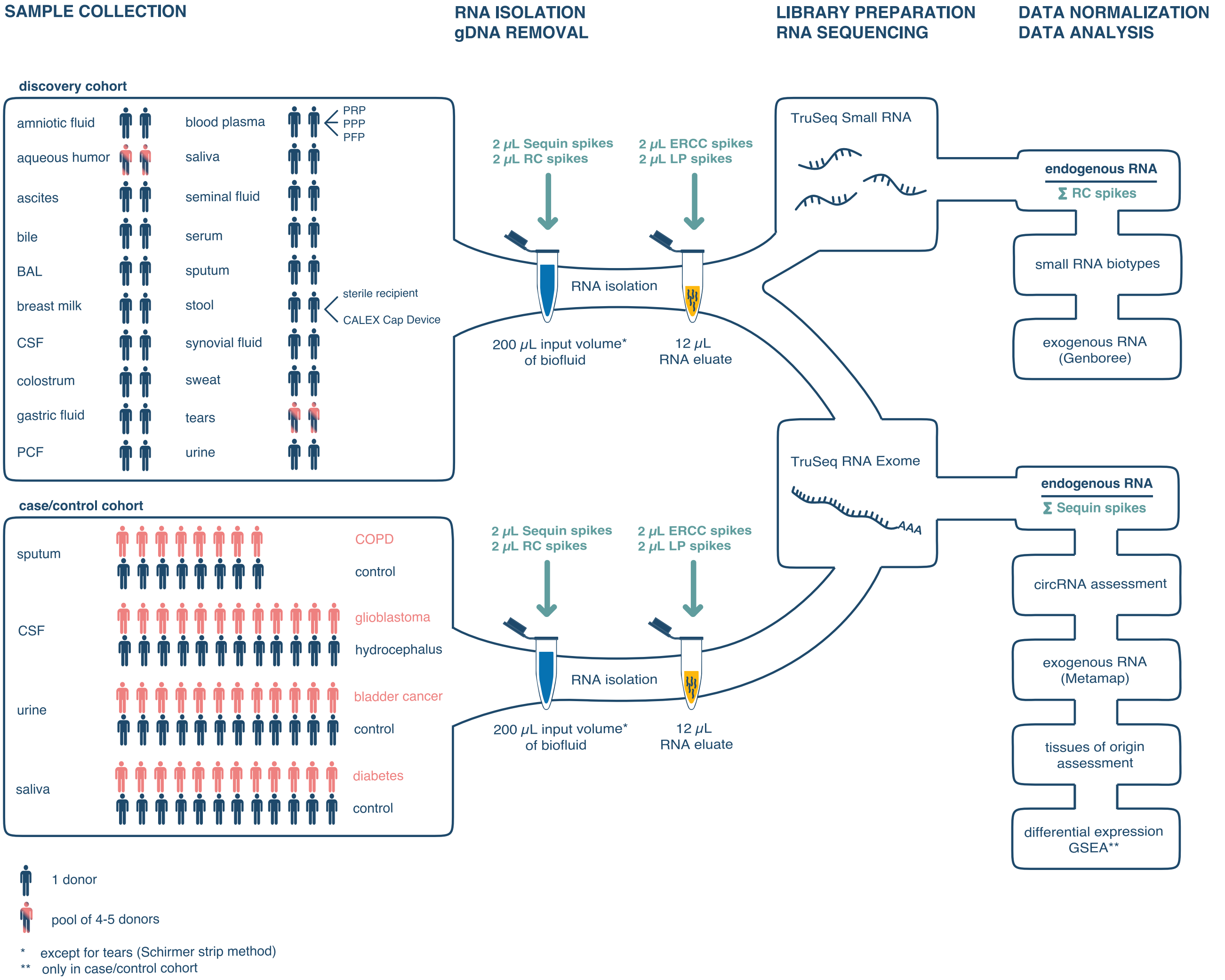


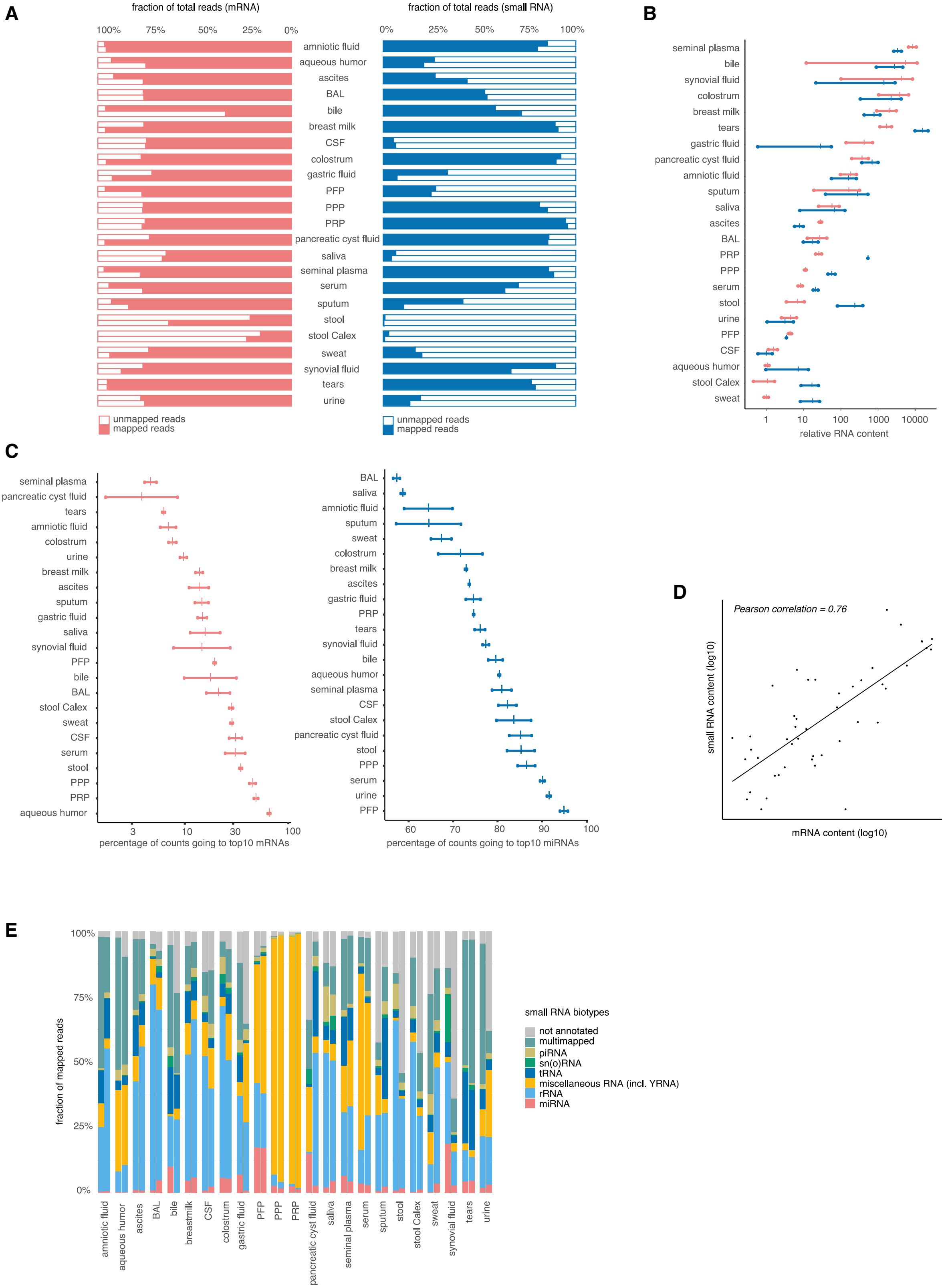
Figure 1. Study Flow Chart

In the discovery cohort, 20 different biofluids were collected in two donors or in a pool of 4–5 donors. In the case/control cohorts, selected biofluids (sputum, CSF, urine, and saliva) were collected in 8–12 patients and an equal number of healthy controls. Both small RNA sequencing and mRNA capture sequencing were performed in the discovery cohort. In the case/control cohorts, mRNA capture sequencing was performed. To compare the RNA content across the different biofluids, the RC spikes and the Sequin spikes are used for normalization of small RNA and mRNA data, respectively. RC and Sequin spikes are added to the biofluid before RNA isolation, and LP and ERCC spikes are added to the RNA eluate. BAL, bronchoalveolar lavage fluid; CSF, cerebrospinal fluid; PRP, platelet-rich plasma; PPP, platelet-poor plasma; PFP, platelet-free plasma.

some biofluids (e.g., seminal plasma and tears), the sequencing coverage of spikes was low, due to a high concentration of endogenous RNA. Detailed information per sample is provided in Figures S1 and S2.

The spike-in controls can also be used to assess the RNA isolation efficiency. The Sequin/ERCC ratio and the RC/LP ratio reflect the relative mRNA and miRNA isolation efficiencies, respectively. A 170-fold difference and a 104-fold difference in RNA isolation efficiency across the samples were observed when assessing long and small RNAs, respectively (Figure S3). These differences underline the challenges of working with heterogenous samples and the importance of spike-in controls for proper data normalization and cross-sample comparison of results.

Finally, the spikes can be utilized to normalize the endogenous RNA abundance data. In this study, we applied a biofluid volume-based normalization by dividing the RNA reads consumed by the endogenous transcripts by the sum of the Sequin spikes for mRNA data and by the sum of the RC spikes for small RNA data. The spike-normalized data represent relative abundance values of RNA molecules proportional to the input volume. Of note, there is an inverse relationship between the number of spike-in RNA reads and the number of endogenous RNA reads. As such, the ratio between the sum of the reads consumed by the endogenous transcripts and the total number of spike-in reads is a relative measure for the RNA concentration of the various samples.



(legend on next page)

Highly Variable mRNA and Small RNA Content among Biofluids in the Discovery Cohort

Both small RNAs and mRNAs were quantified in each of the 20 biofluids in the discovery cohort. Mapping rates varied substantially across the different biofluids (Figure 2A). In general, the proportion of mapped reads was higher for the mRNA-capture-sequencing data (further referred to as mRNA data) than for the small-RNA-sequencing data, in line with the fact that human mRNAs were enriched using biotinylated capture probes during the library preparation. The fraction of mapped reads in the mRNA data ranged from 16% in stool to 97% in seminal plasma. Low mapping rates were observed in stool, in one of the bile samples, and in saliva. Mapping rates for samples in the case/control cohorts are in line with those of the discovery cohort (Figure S4A). In the small-RNA-sequencing data, the proportion of mapped reads ranged from ~7% in stool, saliva, and CSF to 95% in platelet-rich plasma (PRP).

A 10,000-fold difference in mRNA and small RNA concentration was observed between the lowest concentrated fluids (i.e., platelet-free plasma [PFP], urine, and CSF), and the highest concentrated biofluids (i.e., tears, seminal plasma, and bile) (Figure 2B). The absolute mRNA concentration for each biofluid was calculated based on the ERCC spikes and ranged from 0.002 ng/mL in aqueous humor to 33.973 ng/mL in bile (Table 1). The generalizability of the difference in mRNA concentration between highly concentrated biofluids (seminal plasma) and lowly concentrated biofluids (CSF) was confirmed in additional samples (Figure S4B). In the discovery cohort, a 5,547-fold difference in mRNA concentration is observed between seminal plasma and CSF; in independent validation samples, a similarly large 19,851-fold difference in mRNA concentration is observed between both biofluids. In the discovery cohort, the mRNA and miRNA concentrations were significantly correlated across biofluids (Pearson correlation coefficient = 0.76, $p = 8.5e{-}10$; Figure 2D). Normalized abundance levels of exRNAs were significantly correlated between biological replicates within each biofluid (Figure S5). The median Pearson correlation coefficients of the mRNA and the small RNA data were 0.84 and 0.92, respectively. Although the mRNA and miRNA data were well correlated in most biofluids (e.g., tears, colostrum, and saliva), correlation in other biofluids (e.g., bile and pancreatic cyst fluid) was poor. These biofluids are obtained with a more challenging collection method involving echo-endoscopy, impacting the reproducibility of collection and the correlation of the RNA content between biological replicates.

The likelihood of identifying RNA biomarkers in a given biofluid will depend not only on its relative RNA concentration but also on its RNA diversity, here approximated by the fraction of read counts consumed by the top 10 most abundant mRNAs/miRNAs (Figure 2C). In aqueous humor, the top 10 mRNAs represent up to 70% of all reads, indicating that this fluid does not contain a rich mRNA repertoire. In both PRP and platelet-poor plasma (PPP), about 50% of all reads go to the top 10 mRNAs. Although amniotic fluid has a median RNA concentration, this fluid seems to contain a diverse mRNA profile, with only 7% of all reads going to the top 10 mRNAs. When looking into the miRNA data, the top 10 miRNAs represent more than 90% of all reads in PFP, urine, and serum. Bronchoalveolar lavage fluid (BAL) contains the most diverse miRNA repertoire, with 57% of all reads going to the top 10 miRNAs. Similar conclusions with respect to biofluid exRNA diversity can be drawn based on the number of miRNAs/mRNAs representing 50% of the counts (Figure S6). RNA diversity is also reflected by the number of detected exRNAs. The total number of mRNAs and miRNAs detected with at least 4 counts in both samples of the same biofluid ranged from 13,722 mRNAs in pancreatic cyst fluid to 107 mRNAs in aqueous humor and from 231 miRNAs in tears to 18 miRNAs in stool (Table 1). The endogenous RNA mass in nanograms present in 1 mL of each biofluid is also provided in Table 1.

The Distribution of Small RNA Biotypes Varies across the Different Biofluids

The distribution of small RNA biotypes shows distinct patterns among the 20 different biofluids (Figure 2). The exceptionally high percentage of miscellaneous RNAs (mainly, Y-RNAs) observed in blood-derived fluids is in line with the findings of a previous study (El-Mogy et al., 2018) and with the Y-RNA function in platelets. The fraction of reads mapping to miRNAs is lower than 15% in all samples except PFP and one synovial fluid sample. Tears, bile, and amniotic fluid have the highest fraction of tRNA fragments, wherea saliva has the highest fraction of piRNAs. The rRNA fraction is higher than 15% in all samples except tears, aqueous fluid, and the three plasma fractions. The majority of these reads map to the 45S ribosomal RNA transcript. The unannotated read fraction contains mapped reads that could not be attributed to one of the small RNA biotypes. These reads most likely originate from fragmented longer RNAs, such as mRNAs and long non-coding RNAs (Figure S7).

circRNAs Are Enriched in Biofluids compared to Tissues

circRNAs are produced from unspliced RNA through a process called backsplicing, where a downstream 5′ donor binds to an upstream 3′ acceptor. circRNAs are resistant to endogenous exonucleases that target free 5′ or 3′ terminal ends. As a result, circRNAs are highly stable and have extended half-lives

Figure 2. mRNA and Small RNA Content Varies across the 20 Biofluids

(A) Percentage of the total read count mapping to the human transcriptome.

(B) Relative RNA concentration per biofluid; every dot represents the relative RNA concentration in one sample, and every vertical mark indicates the mean per biofluid.

(C) The diversity of the RNA content expressed as fraction of read counts consumed by the top 10 most abundant mRNAs/miRNAs. Only genes with at least 4 unique reads are taken into account. Every dot represents the fraction in one sample, and every vertical mark indicates the mean percentage per biofluid.

(D) Correlation between the small RNA and the mRNA relative concentrations. The Pearson correlation coefficient is 0.76 ($p = 8.5e{-}10^{-10}$). The correlation coefficients are calculated on $\log_{10}$-transformed data.

(E) The fraction of reads that align to small RNA biotypes are shown per biofluid. Only mapped reads of the small-RNA-sequencing data are taken into account. BAL, bronchoalveolar lavage fluid; CSF, cerebrospinal fluid; miRNA, microRNA; PFP, platelet-free plasma; PPP, platelet-poor plasma; PRP, platelet-rich plasma; piRNAs, piwi-interacting RNAs; snRNAs, small nuclear RNAs; snoRNAs, small nucleolar RNAs; tRNAs, transfer RNA.

Table 1. Endogenous mRNA Concentration and Number of RNAs per Biofluid

Biofluid	Endogenous mRNA Concentration (ng/mL)	Number of miRNAs	Number of mRNAs
Bile	33.973	45	2,279
Seminal plasma	22.351	211	11,868
Synovial fluid	17.595	122	1,614
Colostrum	13.535	229	11,914
Breast milk	7.463	213	11,607
Tears	7.316	231	13,366
Pancreatic cyst fluid	1.643	129	13,722
Sputum	0.297	91	7,738
Amniotic fluid	0.206	119	10,531
Gastric fluid	0.181	21	9,288
Saliva	0.171	110	6,353
PRP	0.073	192	5,440
Ascites	0.056	75	5,578
BAL	0.053	126	3,565
PPP	0.026	113	4,548
Serum	0.023	122	4,152
Urine	0.016	41	2,094
PFP	0.013	95	2,699
Stool Calex	0.005	18	135
Stool	0.005	19	134
CSF	0.005	32	438
Sweat	0.002	45	410
Aqueous humor	0.002	20	107

For each biofluid, the mean mass of endogenous RNAs in nanograms detected per 1 mL biofluid is provided. The number of mRNAs and miRNAs with at least 4 unique read counts in both replicates is shown per biofluid. BAL, bronchoalveolar lavage fluid; CSF, cerebrospinal fluid; PFP, platelet-free plasma; PPP, platelet-poor plasma; PRP, platelet-rich plasma.

compared to linear mRNAs (Li et al., 2018b). circRNAs have been reported to be present in numerous human tissues (Vo et al., 2019) and in a few biofluids, such as saliva (Bahn et al., 2015), blood (Memczak et al., 2015), semen (Liu et al., 2019) and urine (Kölling et al., 2019; Vo et al., 2019). A direct comparison of the circRNA read fraction between biofluids and tissues is currently lacking in literature. We compared the circRNA fraction, for genes that produce both linear and circular transcripts, identified through mRNA capture sequencing of the 20 biofluids in this study, with the circRNA fraction identified in mRNA capture sequencing of 36 cancerous tissue types obtained from the MiOncoCirc Database (Vo et al., 2019). Although more unique backsplice junctions were identified in tissues compared to biofluids, in line with the higher RNA concentration in tissues (Figure 3B), the circRNA read fraction is clearly higher in biofluid exRNA compared to cellular RNA (Figure 3A). The median circRNA read fraction in biofluids is 84.4%, which is significantly higher than the median circRNA read fraction in tissues of 17.5% (Mann-Whitney U test, two-sided, p = 5.36e−12). For genes that produce both linear and circular transcripts, the stable circRNAs are more abundant than the linear mRNAs in biofluids, whereas it is the other way around in tissues.

We used two different methods to define the circRNA read fraction (see "Circular RNA detection and circular/linear ratio determination" in STAR Methods; Figure S8): one based on individual backsplice junctions (shown in Figure 3) and another method based on backsplice junctions aggregated at gene level (Figure S9). Both methods clearly point toward a substantial enrichment of circRNAs in biofluids.

Assessment of Exogenous RNA in Human Biofluids

Two dedicated pipelines were used for the non-trivial assessment of the presence of microbial or viral RNA in human biofluid exRNA. Overall, the fraction of bacterial reads is higher in small-RNA-sequencing data than in the mRNA data, in line with the unbiased nature of small RNA sequencing and the targeted hybrid capture enrichment using probes against human RNA during the mRNA capture library preparation (Figure 4A). Stool (both collection methods), sweat, saliva, and sputum are among the biofluids with the highest fraction of bacterial RNA in both the small-RNA-sequencing data and the mRNA data. The percentage of bacterial reads in mRNA data and in small RNA data are significantly correlated across biofluids (Pearson correlation coefficient = 0.78, p = 1.94e−10).

Bacterial reads in aqueous humor and CSF, two fluids with very low endogenous RNA content that were collected in a sterile setting (and, thus, presumed to be sterile), most likely reflect background contamination during the workflow (Heintz-Buschart et al., 2018). To illustrate the biological relevance of the bacterial signal, we looked into reads mapping to *Campylobacter concisus*, a Gram-negative bacterium that is known to primarily colonize the human oral cavity, with some strains translocated to the intestinal tract (Liu et al., 2018). We confirm the selective presence of reads mapping to *Campylobacter concisus* in saliva in both the small RNA and the mRNA data (Figure 4B). In all samples and for both the small RNA and the mRNA data, the percentage of the total reads that maps to viral transcriptomes is less than 1%.

Assessment of the Tissues of Origin and Deconvolution of Pancreatic Cyst Fluid

Gaining insights in tissue contribution to biofluid RNA profiles may guide the selection of the most appropriate biofluid to investigate a given disease. To define tissues that specifically contribute RNA molecules to individual biofluids, we explored the relationship between extracellular mRNA levels and tissue- or cell-type-specific mRNA signatures. The heatmap in Figure 5A highlights the relative contribution of tissues and cell types to a specific biofluid compared to the other biofluids. More detailed results per biofluid are shown in Figure S10. The results of this analysis were validated in an independent sample cohort for CSF, saliva, sputum, seminal plasma, and urine (Figure S4C). As expected, prostate tissue RNA markers are more abundant in urine and in seminal plasma than in any other biofluid. Both sputum and saliva contain mRNAs specific for trachea and esophagus. In amniotic fluid, markers for esophagus, small intestine, colon, and lung are more abundant than for the other tissues and cell types, probably reflecting

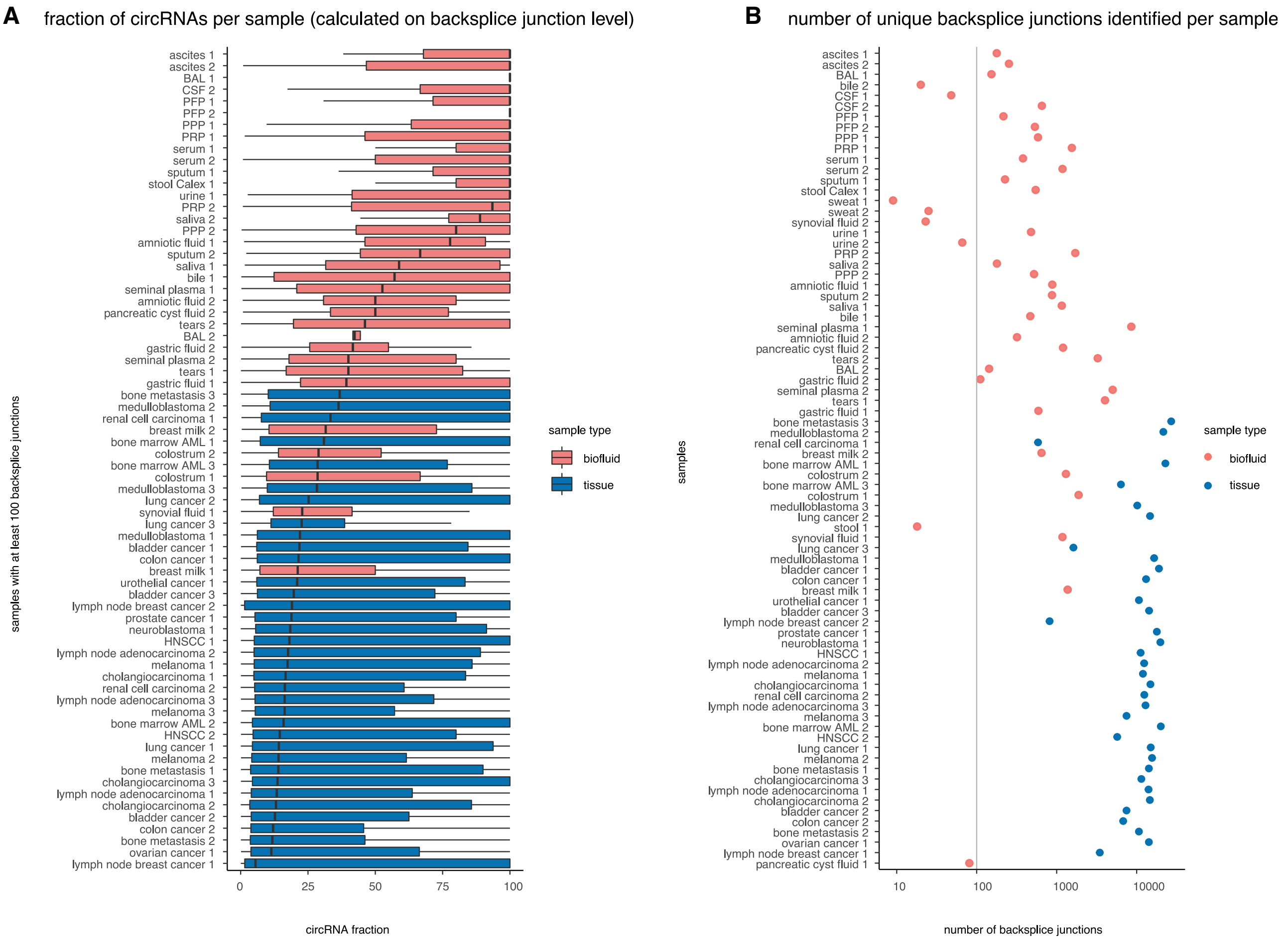


Figure 3. circRNAs Are Enriched in Biofluids compared to Tissues

(A) The circRNA fraction, calculated at the backsplice junction level, is plotted per sample and is higher in cell-free biofluid RNA than in tissue RNA. Only samples with at least 100 backsplice junctions are plotted.

(B) The number of unique backsplice junctions per sample is higher in tissues compared to biofluids, in line with the higher input concentration of RNA into the library prep.

AML, acute myeloid leukemia; BAL, bronchoalveolar lavage fluid; CSF, cerebrospinal fluid; HNSCC: head and neck squamous cell carcinoma; PFP, platelet-free plasma; PPP, platelet-poor plasma; PRP, platelet-rich plasma.

organs that actively shed RNA (at the gestational age of sampling) into the amniotic cavity. These data strongly suggest that biofluid mRNA levels, at least to some degree, reflect intracellular mRNA levels from cells that produce or transport the fluid. To further investigate the origin of biofluid RNA at the cellular level, we applied computational deconvolution of the pancreatic cyst fluid RNA profiles using single-cell RNA-sequencing data from 10 pancreatic cell types (Baron et al., 2016). Figure 5B reveals that pancreatic cyst fluid 1 consists of 45% activated stellate cells and 43% endothelial cells, while pancreatic cyst fluid 2 mainly consists of quiescent stellate cells (38%), endothelial cells (31%), and acinar cells (19%).

Biomarker Potential of mRNA in Sputum, Urine, CSF, and Saliva in Selected Case/Control Cohorts

Additional biofluid samples were collected in patients with a specific disease or in healthy controls to investigate potential biologically relevant differences in mRNA content between both groups. Sequin RNA spikes were used for biofluid volume-based data normalization. Strikingly, the relative RNA concentration in sputum of COPD patients was higher than in non-COPD patients, probably reflecting the high turnover of immune cells during the state of chronic inflammation (Figure 6A). Differential expression analysis revealed 5,513 and 6 mRNAs that were significantly up- and downregulated, respectively, in sputum from COPD patients compared to that from healthy controls (Figure 6B). CCL20, the most differential mRNA, showed a 146-fold upregulation in COPD patients compared to that in healthy donors. This potent chemokine attracting dendritic cells has previously been linked to the pathogenesis of COPD (Bracke et al., 2006; Demedts et al., 2007); ADA and MMP1, also among the most differential mRNAs, have also been associated with the pathogenesis of COPD (Karmouty-Quintana et al., 2013; Singh Patidar et al., 2018; Stankovic et al., 2017). To verify the

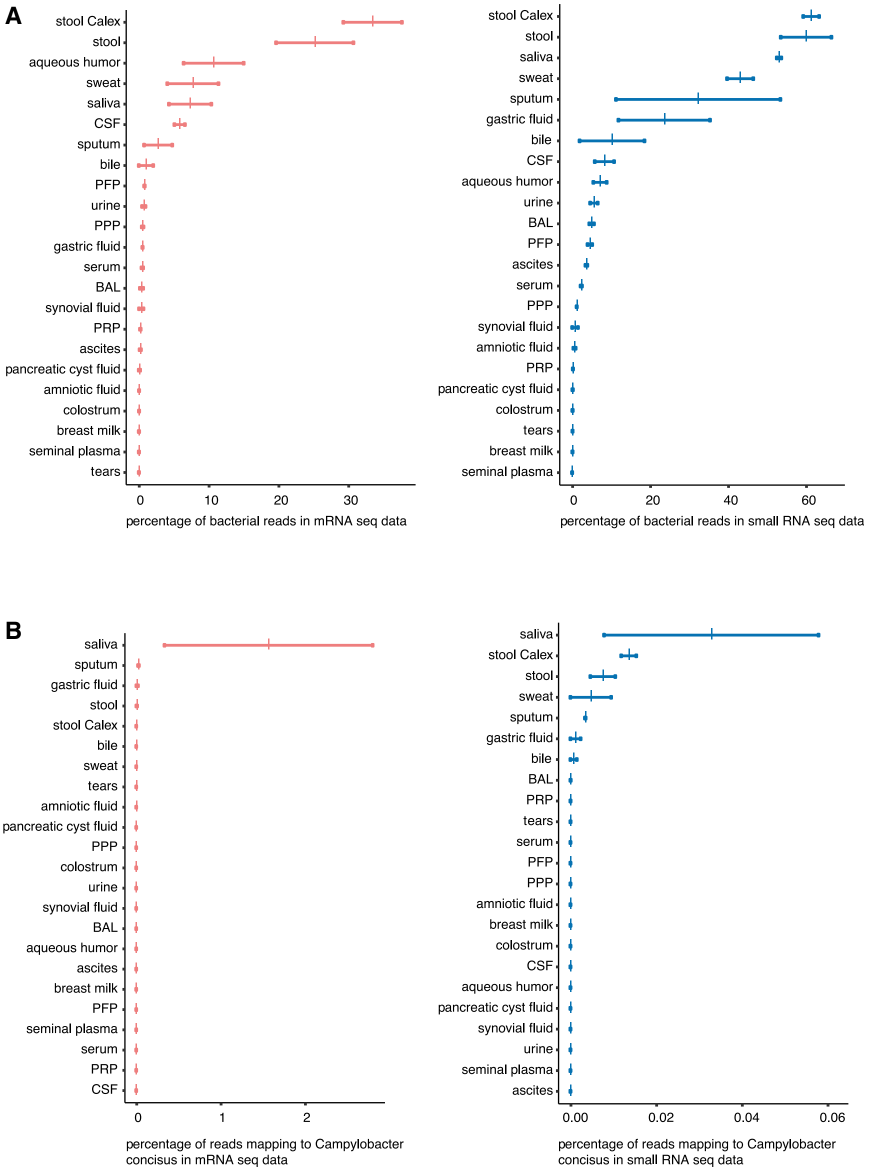


Figure 4. Reads Mapping to Bacterial Genomes

(A) Percentage of reads mapping to bacteria in mRNA data (pink) and in small-RNA-sequencing data (blue).

(B) Percentage of reads mapping to *Campylobacter concisus* in mRNA data (pink) and in small-RNA-sequencing data (blue). *Campylobacter concisus* is known to be present in saliva.

RNA-sequencing findings, 8/8 of the most differentially abundant mRNAs were validated by qRT-PCR (Figures S11A and S11B).

In contrast to that in patients with COPD, the relative RNA content is comparable in urine from bladder cancer patients and healthy volunteers, in CSF from glioblastoma patients and hydrocephalus patients, and in saliva from diabetes patients and healthy volunteers (Figures 6C–6E and S12). A higher RNA yield in CSF from glioblastoma patients compared to that in CSF from healthy controls has been reported by Saugstad et al. (2017); however, the collection method of CSF differed between both groups, and it is therefore not possible to assess whether the reported difference in RNA yield between both groups is due to the different CSF collection sites (lumbar puncture versus craniotomy) or due to the neurological disease. In urine from patients with a muscle-invaded bladder cancer, 529 mRNAs and 9 mRNAs were significantly upregulated and downregulated, respectively, compared to urine from healthy volunteers (Figure 6D). Some of the upregulated mRNAs, such as MDK, SLC2A1, GPRC5A, KRT17, and KRT5, have been reported in urine and were suggested as biomarkers for the accurate detection and classification of bladder cancer (Eckstein et al., 2018; Holyoake et al., 2008; Lin et al., 2019; Murakami et al., 2018). In CSF from glioblastoma patients, only 2 mRNAs are significantly upregulated compared to that in CSF from hydrocephalus patients. CD163, one of the upregulated genes in glioblastoma, has been linked with glioblastoma pathogenesis (Chen et al., 2019). In saliva from diabetes patients and saliva from healthy volunteers, no differentially expressed genes could be identified.

Differential abundance analysis was performed for circRNAs as well, but in none of the case/control cohorts could differentially abundant circRNAs be detected (data not shown). As circRNAs can only be identified based on their backsplice junction, the read coverage is generally (too) low for biomarker discovery based on mRNA-capture-sequencing data. When applying a similar strategy for mRNAs by looking at the reads of only one "linear-only" junction per gene (outside every detected backsplice junction) a significantly lower number of differentially abundant mRNAs was detected (sputum: 13 out of 5,519 mRNAs; urine: 0 out of 538 mRNAs; CSF: 0 out of 35 mRNAs). These results strongly suggest that a dedicated circRNA enrichment strategy may be needed to assess circRNA biomarker potential.

To validate the identification of the 10 most abundant circRNAs detected by mRNA capture sequencing in sputum, an orthogonal validation by qRT-PCR of the backsplice sequence region was performed. For 9 of the 10 circRNAs, the RNA-sequencing results could be validated (Figure S11C).

DISCUSSION

By applying two complementary RNA-sequencing technologies on 20 different biofluids, we assembled the most comprehensive

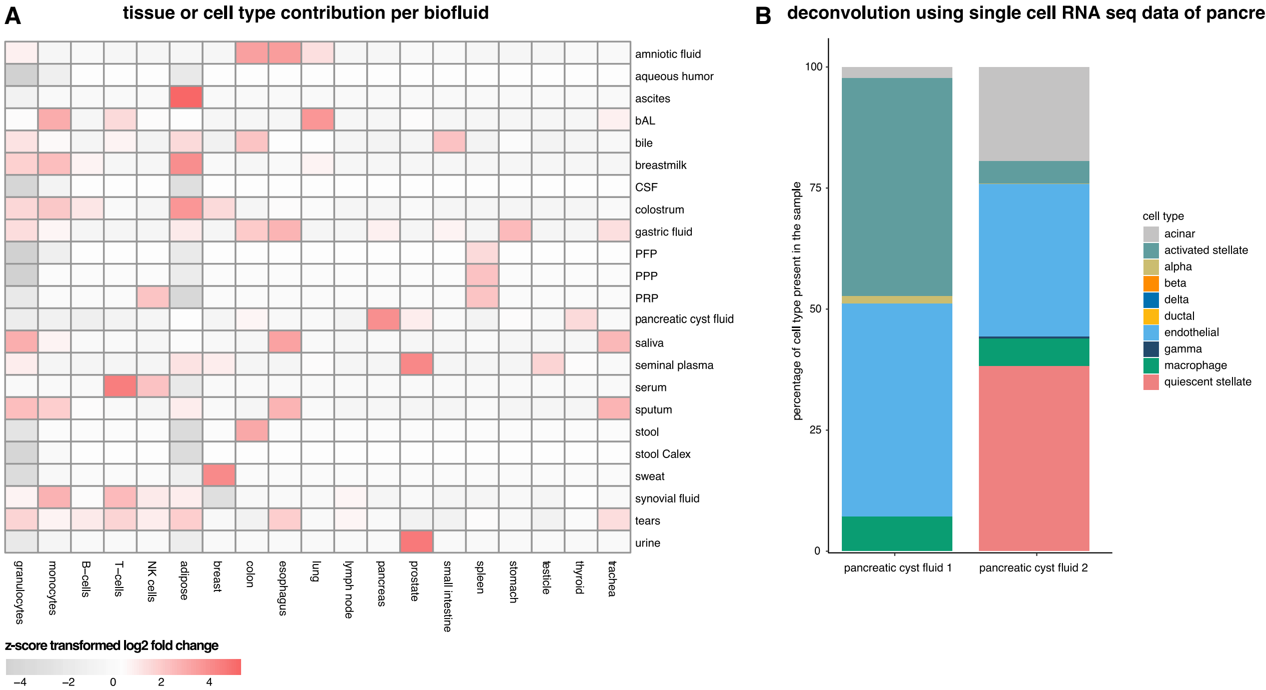


Figure 5. Identification of the Tissues of Origin per Biofluid and Deconvolution of Pancreatic Cyst Fluid

(A) Assessment of the tissues of origin in the biofluids of the discovery cohort.

(B) Heatmap showing tissues and cell types that contribute more specifically to a certain biofluid compared to the other biofluids. Rows depict the biofluids of the discovery cohort, and the columns indicate the tissues or cell types for which markers were selected based on the RNA Atlas (Lorenzi et al., 2019). For visualization purposes, only tissues and cell types with a Z-score-transformed $\log_2$ fold change $\geq |1|$ in at least one biofluid are indicated.

(C) Composition of pancreatic cyst fluid samples based on deconvolution using sequencing data from 10 pancreatic cell types.

human biofluid transcriptome, covering small RNAs, mRNAs and circRNAs. Until now, most efforts to investigate and compare the RNA content within biofluids focused on small RNA sequencing, most likely because of technical limitations and unawareness of the abundance of extracellular mRNA (fragments) (El-Mogy et al., 2018; Ferrero et al., 2017; Godoy et al., 2018; Srinivasan et al., 2019; Weber, 2017; Yeri et al., 2017).

The availability of both small RNA-sequencing data and mRNA data allows a more in-depth characterization of the human transcriptome in biofluids. To our knowledge, this is the first study reporting on the mRNA content, generated through a dedicated mRNA enrichment-sequencing method, in tear fluid, amniotic fluid, aqueous humor, bile, bronchial lavage fluid, gastric fluid, saliva, seminal plasma, synovial fluid, sweat, and urine. Selected mRNAs were previously studied by means of qRT-PCR in amniotic fluid (Welch et al., 2016), pancreatic cyst fluid (Maker et al., 2019; Marzioni et al., 2015), seminal plasma (Tian et al., 2016), sputum (Oreo et al., 2014), stool (Herring et al., 2018), and in extracellular vesicles isolated from cell-free urine (Bazzell et al., 2018). In saliva, selected mRNAs were detected using microarrays (Zhang et al., 2010). We have demonstrated that it is technically feasible to generate mRNA data from low-input biofluid samples. This is expected to accelerate biomarker research in these fluids. Further efforts to profile and share the mRNA and circRNA content in larger sample cohorts of biofluids, comparable to the exRNA Atlas Resource for small RNAs, are necessary to move this scientific field forward (Murillo et al., 2019).

Synthetic spike-in controls allowed for a direct comparison of RNA content across biofluids, revealing a 10,000-fold difference in concentration. Of note, the RNA concentration is not perfectly correlated with the transcriptome complexity (as reflected by the number of miRNAs and the number of mRNAs detected per sample; Table 1).

Our small RNA results confirm previous studies observing a high miRNA concentration in tears (Weber et al., 2010), low mapping rates in CSF (Godoy et al., 2018; Waller et al., 2018), and a low miRNA concentration in cell-free urine (El-Mogy et al., 2018). A direct comparison of the absolute numbers of detected miRNAs, mRNAs, and circRNAs detected per sample in our study with the numbers in published literature is hampered by the fact that the absolute read count is dependent on the input volume of the biofluids, the RNA isolation kit, and library preparation method used, the sequencing depth, and data-analysis settings (e.g., mapping without mismatches and filtering of the data). In addition, different pre-analytical variables when preparing the biofluid samples may also affect the sequencing results. However, on a higher level, we can look into the most abundant miRNAs detected in specific biofluids. The majority of the 10 most abundant miRNAs detected in 9 specific biofluids reported by Godoy et al. (2018) are also detected among the most abundant miRNAs in the samples from the discovery cohort (Table S1).

We compared the mRNA results of the discovery cohort with these of the case/control cohorts. Mapping rates for samples in the discovery cohort are in the same range for saliva, sputum,

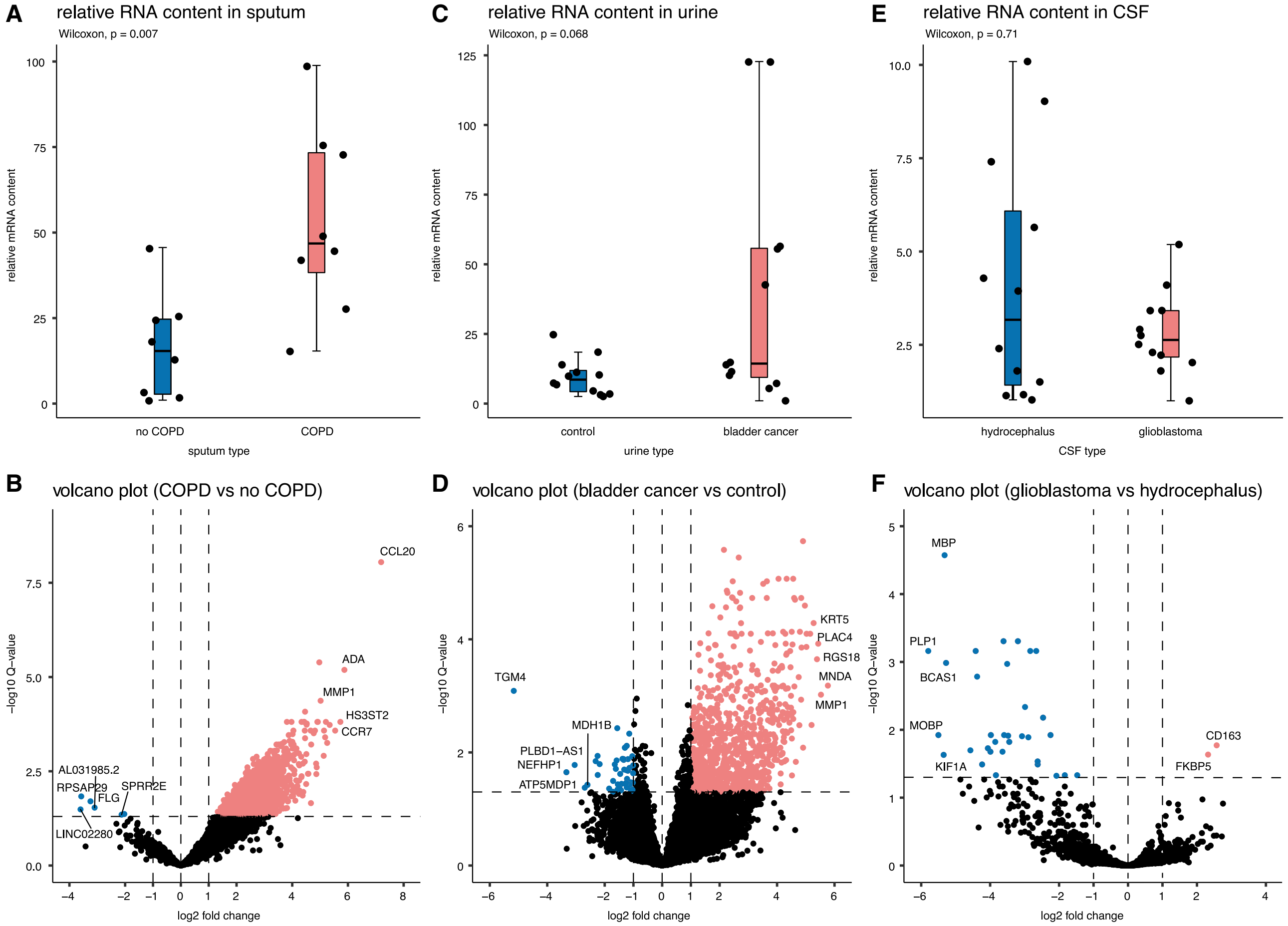


Figure 6. Relative RNA Concentration and Volcano Plot in Case/Control Cohorts

Boxplots of relative mRNA content (top) and volcano plots of differentially expressed mRNAs (bottom) ($q < 0.05$; pink indicates up, and blue indicates down in patient versus control) with labeling of up to the 5 most differential genes.

(A) Sputum from COPD patients (n = 8) compared to sputum from healthy donors (n = 8; Wilcoxon rank test, two-sided, $p = 7e{-}3$).

(B) 5,513 and 6 mRNAs up and down, respectively, in COPD samples.

(C) Urine from bladder cancer patients (n = 12) compared to urine from healthy donors (n = 12; Wilcoxon signed-rank test, two-sided, $p = 6.8e{-}2$).

(D) 529 and 9 mRNAs up and down, respectively, in bladder cancer samples.

(E) CSF from glioblastoma cancer patients (n = 12) compared to CSF from hydrocephalus patients (n = 12; Wilcoxon signed-rank test, two-sided, $p = 7.1e{-}1$).

(F) 2 and 33 mRNAs up and down, respectively, in glioblastoma samples.

and seminal plasma. The mapping rates for CSF and urine are about 15% higher in the case/control cohorts compared to those in the discovery cohort. These differences may be due to different pre-analytical variables between both cohorts (collection tube, centrifugation speed, and the portion of urine collected) (Figure S4A; Methods S1). Note that sputum in the discovery cohort was collected as spit samples from volunteers with a common cold, while all 16 sputum samples of the case/control cohort were collected through sputum induction.

In the discovery cohort, on average, 53% of all small RNA reads in saliva can be traced to bacteria, perfectly in line with the average of 45.5% reads mapping to bacteria reported by Yeri et al. (2017). Aqueous humor and CSF, although collected in a sterile setting and presumed to be sterile, contain up to 11% of reads mapping to bacteria, in line with previous studies (Godoy et al., 2018; Waller et al., 2018). However, bacterial cultures of our two CSF samples were negative. As both CSF and aqueous humor display a very low relative RNA content, the exogenous sequences may represent bacterial contaminants introduced during the sample processing workflow. Contaminants can derive from contaminated spin columns used during RNA purification (Heintz-Buschart et al., 2018), enzymes produced in microorganisms (Salter et al., 2014), or various environmental sources (Strong et al., 2014). Such contaminant signals are likely underrepresented in samples with a high concentration of endogenous exRNAs. The exogenous RNA content was assessed in both the mRNA-capture-sequencing data and the small-RNA-sequencing data. As the mRNA capture probes were not designed to capture exogenous RNA, mRNA capture sequencing is not the preferred method for bacterial RNA quantification. Despite the shortcomings of the capture technique, the abundance estimates of the bacterial species derived from the mRNA capture data match with those derived from the small-RNA-sequencing data (the latter being unbiased with respect to bacterial RNA quantification).

Although we collected a broad range of biofluids, only two samples per biofluid were studied, limiting our ability to assess donor variability. The input volume for the RNA isolations in all biofluids was set to 200 μL, and a volume-based comparison of the RNA content was made among the biofluids. We did not explore whether higher input volumes would result in higher RNA yields in biofluids where this could have been possible (e.g., urine). We also note that the results in Table 1 are impacted by biofluid input volume in the RNA purification, RNA input in the sequencing library preparation, and the sequencing depth.

Biofluid data normalization with synthetic spike-in controls is a unique and powerful approach and reflects more accurately the biological situation compared to classic normalization approaches where global differences on overall abundance are neutralized. For instance, the relative mRNA concentration in sputum from COPD patients is higher than in sputum from healthy donors. Typically, RNA-sequencing data are subsampled or normalized based on the library size before performing a differential expression analysis, resulting in an artificially more balanced volcano plot, an overcorrection of the biological situation, and a loss of information, which is not the case when the data are normalized based on spike-in controls.

Our results highlighting tissues and cell types that contribute more specifically to a certain biofluid compared to the other biofluids (Figure 5A) can be used as a roadmap to formulate hypotheses when initiating biomarker research. Not surprisingly, the RNA signal from prostate is reflected in urine and seminal plasma. Both fluids can be collected in a non-invasive way and may be of value to investigate further in prostate cancer patients. Of interest, the mRNA concentration in seminal plasma is 1,000-fold higher than in urine, and seminal plasma contains more unique mRNAs compared to urine, suggesting that the biomarker potential of seminal plasma is higher. However, one should also be cautious in interpreting the tissue enrichment results: although the RNA signal of breast seems relatively enriched in sweat, this biofluid has the lowest RNA concentration. The limited number of detected mRNAs in sweat show overlap with mRNAs related to secretion (MCL1 gene, SCGB2A2 gene, and SCGB1D2 gene) that also appear as markers in breast tissue.

The pancreatic tissue RNA signal appears to be enriched in pancreatic cyst fluid, and a different cell-type composition is observed when both samples are deconvoluted using single-cell RNA-sequencing data of pancreatic cell types (Figure 5B). Pancreatic cyst fluid was collected in these donors to investigate a cystic lesion in the pancreas. The routine cytological analysis of these fluid samples was inconclusive at the moment of sample collection. By following up both patients, we discovered that the first patient developed a walled-off necrosis collection after necrotizing pancreatitis. The incipient high fraction of activated stellate cells in the first cyst fluid sample may have been an indication pointing toward the inflammation and necrosis that finally occurred. The second patient was diagnosed with a side-branch intra-papillary mucinous neoplasia, probably reflected by the relative high fraction of acinar cells. Pancreatic cysts are often detected on abdominal imaging, resulting in a diagnostic and treatment dilemma. Furthermore, pancreatic cysts represent a broad group of lesions, ranging from benign to malignant entities. The main challenge in their management is to accurately predict the malignant potential and to determine the risk to benefit of a surgical resection (Farrell, 2017). Our results show that the cellular contribution to the RNA content of pancreatic cyst fluids can be estimated through deconvolution and that these results may be associated with clinical phenotypes. Larger cohorts are necessary to investigate the clinical potential of this approach, and pancreatic tumor cells may also need to be added to the reference set with single-cell RNA-sequencing data to improve the accuracy of the prediction.

In addition to linear mRNA transcripts, we also explored the circRNA content in biofluids. circRNAs are a growing class of non-coding RNAs and a promising RNA biotype to investigate in the liquid biopsy setting, as they are presumed to be less prone to degradation compared to linear forms (Jeck and Sharpless, 2014). circRNAs can be detected with mRNA capture sequencing through the capture of exons that are incorporated in the circRNAs, followed by identification of the characteristic backsplice junction. The circRNA fraction in tissues has previously been reported and is in line with our findings (Guo et al., 2014). In our study, we demonstrated that, for genes that produce both circRNAs and linear mRNAs, the circRNAs are more abundant than the linear forms in biofluids. Further assessment of the biomarker potential of circRNAs in biofluids require dedicated library preparation methods with circRNA enrichment.

In conclusion, The Human Biofluid RNA Atlas provides a systematic and comprehensive comparison of the exRNA content in 20 different human biofluids. The results presented here may serve as a valuable resource for future biomarker studies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead Contact
  - ○ Materials Availability
  - ○ Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Donor material, collection and biofluid preparation procedure
- METHOD DETAILS
  - ○ RNA isolation and gDNA removal
  - ○ TruSeq RNA Exome library prep sequencing
  - ○ TruSeq Small RNA library prep sequencing
  - ○ RT-qPCR
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Processing TruSeq RNA Exome sequencing data
  - ○ Processing TruSeq Small RNA sequencing data
  - ○ Calculation of endogenous RNA concentration
  - ○ Exogenous RNA characterization
  - ○ Circular RNA detection and circular/linear ratio determination

- ○ Assessment of tissue and cell contribution to biofluid exRNA
- ○ Cellular deconvolution of pancreatic cyst fluid samples
- ○ Differential expression analysis in case/control cohorts

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2020.108552.

ACKNOWLEDGMENTS

E.H. is a recipient of a grant from the Fund for Scientific Research Flanders (FWO; 1133120N). F.A.C. and A.M. are supported by Special Research Fund (BOF) scholarships of Ghent University (BOF.DOC.2017.0026.01 and BOF.DOC.2019.0047.01). A.G. is senior clinical researcher of the FWO (FWO; 1805718N). This research is partly funded by "RNA-MAGIC" and "LNCCA" Concerted Research Actions of Ghent University (BOF19/GOA/008 and BOF16/GOA/023), by "Kom Op Tegen Kanker" (Stand Up To Cancer, the Flemish cancer society), and by Czech Ministry of Health grants (NV18-03-00398 and NV18-03-00360). We thank Tim Mercer for providing the Sequin spikes.

The results published here are, in part, based on data generated by the Clinical Sequencing Exploratory Research (CSER) consortium established by the NHGRI. Funding support was provided through cooperative agreements with the NHGRI and NCI through grant number U01 HG006508 (Exploring Cancer Medicine for Sarcoma and Rare Cancers). Information about CSER and the investigators and institutions who constitute the CSER consortium can be found at https://www.genome.gov/27546194.

AUTHOR CONTRIBUTIONS

J.V. and P.M. conceived and supervised the project; E.H., K.V., N.Y., E.V.E., and J.N. designed and performed the experiments; E.H., A.M., J.A., and F.A.C. analyzed the data; L.S. and S. Karg performed analysis using the MetaMap pipeline; G.P.S. and S. Kuersten contributed technical support and resources; E.H., A.G., P.H., P.J., G.B., K.R.B., T. Maes, T. Malfait, T.D., V.N., C.V.C., K.R., E.R., D.H., K.T., O.S., and C.N. collected samples; S.L. designed qRT-PCR primers; E.H., P.M., and J.V. wrote the paper; J.K. developed dedicated tools to analyze RNA atlas data and results and implemented them in the online portal R2. All authors contributed to manuscript editing and approved the final draft.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 18, 2020
Revised: October 14, 2020
Accepted: December 3, 2020
Published: December 29, 2020

REFERENCES

Avila Cobos, F., Vandesompele, J., Mestdagh, P., and De Preter, K. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics *34*, 1969–1979.

Bahn, J.H., Zhang, Q., Li, F., Chan, T.-M., Lin, X., Kim, Y., Wong, D.T.W., and Xiao, X. (2015). The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva. Clin. Chem. *61*, 221–230.

Baron, M., Veres, A., Wolock, S.L., Faust, A.L., Gaujoux, R., Vetere, A., Ryu, J.H., Wagner, B.K., Shen-Orr, S.S., Klein, A.M., et al. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst. *3*, 346–360.e4.

Bazzell, B.G., Rainey, W.E., Auchus, R.J., Zocco, D., Bruttini, M., Hummel, S.L., and Byrd, J.B. (2018). Human Urinary mRNA as a Biomarker of Cardiovascular Disease. Circ. Genom. Precis. Med. *11*, e002213.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B *57*, 289–300.

Bracke, K.R., D'hulst, A.I., Maes, T., Moerloose, K.B., Demedts, I.K., Lebecque, S., Joos, G.F., and Brusselle, G.G. (2006). Cigarette smoke-induced pulmonary inflammation and emphysema are attenuated in CCR6-deficient mice. J. Immunol. *177*, 4350–4359.

Chen, T., Chen, J., Zhu, Y., Li, Y., Wang, Y., Chen, H., Wang, J., Li, X., Liu, Y., Li, B., et al. (2019). CD163, a novel therapeutic target, regulates the proliferation and stemness of glioma cells via casein kinase 2. Oncogene *38*, 1183–1199.

Demedts, I.K., Bracke, K.R., Van Pottelberge, G., Testelmans, D., Verleden, G.M., Vermassen, F.E., Joos, G.F., and Brusselle, G.G. (2007). Accumulation of dendritic cells and increased CCL20 levels in the airways of patients with chronic obstructive pulmonary disease. Am. J. Respir. Crit. Care Med. *175*, 998–1005.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Eckstein, M., Wirtz, R.M., Gross-Weege, M., Breyer, J., Otto, W., Stoehr, R., Sikic, D., Keck, B., Eidt, S., Burger, M., et al. (2018). mRNA-Expression of *KRT5* and *KRT20* Defines Distinct Prognostic Subgroups of Muscle-Invasive Urothelial Bladder Cancer Correlating with Histological Variants. Int. J. Mol. Sci. *19*, 3396.

El-Mogy, M., Lam, B., Haj-Ahmad, T.A., McGowan, S., Yu, D., Nosal, L., Rghei, N., Roberts, P., and Haj-Ahmad, Y. (2018). Diversity and signature of small RNA in different bodily fluids using next generation sequencing. BMC Genomics *19*, 408.

Everaert, C., Helsmoortel, H., Decock, A., Hulstaert, E., Van Paemel, R., Verniers, K., Nuytens, J., Anckaert, J., Nijs, N., Tulkens, J., et al. (2019). Performance assessment of total RNA sequencing of human biofluids and extracellular vesicles. Sci. Rep. *9*, 17574.

Farrell, J.J. (2017). Pancreatic Cysts and Guidelines. Dig. Dis. Sci. *62*, 1827–1839.

Fehlmann, T., Ludwig, N., Backes, C., Meese, E., and Keller, A. (2016). Distribution of microRNA biomarker candidates in solid tissues and body fluids. RNA Biol. *13*, 1084–1088.

Ferrero, G., Cordero, F., Tarallo, S., Arigoni, M., Riccardo, F., Gallo, G., Ronco, G., Allasia, M., Kulkarni, N., Matullo, G., et al. (2017). Small non-coding RNA profiling in human biofluids and surrogate tissues from healthy individuals: description of the diverse and most represented species. Oncotarget *9*, 3097–3111.

Freedman, J.E., Gerstein, M., Mick, E., Rozowsky, J., Levy, D., Kitchen, R., Das, S., Shah, R., Danielson, K., Beaulieu, L., et al. (2016). Diverse human extracellular RNAs are widely detected in human plasma. Nat. Commun. *7*, 11106.

Giraldez, M.D., Spengler, R.M., Etheridge, A., Goicochea, A.J., Tuck, M., Choi, S.W., Galas, D.J., and Tewari, M. (2019). Phospho-RNA-seq: a modified small RNA-seq method that reveals circulating mRNA and lncRNA fragments as potential biomarkers in human plasma. EMBO J. *38*, e101695.

Godoy, P.M., Bhakta, N.R., Barczak, A.J., Cakmak, H., Fisher, S., MacKenzie, T.C., Patel, T., Price, R.W., Smith, J.F., Woodruff, P.G., and Erle, D.J. (2018). Large Differences in Small RNA Composition Between Human Biofluids. Cell Rep. *25*, 1346–1358.

Green-Church, K.B., Nichols, K.K., Kleinholz, N.M., Zhang, L., and Nichols, J.J. (2008). Investigation of the human tear film proteome using multiple proteomic approaches. Mol. Vis. *14*, 456–470.

Guo, J.U., Agarwal, V., Guo, H., and Bartel, D.P. (2014). Expanded identification and characterization of mammalian circular RNAs. Genome Biol. *15*, 409.

Hafner, M., Renwick, N., Brown, M., Mihailović, A., Holoch, D., Lin, C., Pena, J.T.G., Nusbaum, J.D., Morozov, P., Ludwig, J., et al. (2011). RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA *17*, 1697–1712.

Heintz-Buschart, A., Yusuf, D., Kaysen, A., Etheridge, A., Fritz, J.V., May, P., de Beaufort, C., Upadhyaya, B.B., Ghosal, A., Galas, D.J., and Wilmes, P. (2018). Small RNA profiling of low biomass samples: identification and removal of contaminants. BMC Biol. *16*, 52.

Herring, E., Kanaoka, S., Tremblay, E., and Beaulieu, J.-F. (2018). A Stool Multitarget mRNA Assay for the Detection of Colorectal Neoplasms. Methods Mol. Biol. *1765*, 217–227.

Holyoake, A., O'Sullivan, P., Pollock, R., Best, T., Watanabe, J., Kajita, Y., Matsui, Y., Ito, M., Nishiyama, H., Kerr, N., et al. (2008). Development of a multiplex RNA urine test for the detection and stratification of transitional cell carcinoma of the bladder. Clin. Cancer Res. *14*, 742–749.

Jeck, W.R., and Sharpless, N.E. (2014). Detecting and characterizing circular RNAs. Nat. Biotechnol. *32*, 453–461.

Karmouty-Quintana, H., Weng, T., Garcia-Morales, L.J., Chen, N.-Y., Pedroza, M., Zhong, H., Molina, J.G., Bunge, R., Bruckner, B.A., Xia, Y., et al. (2013). Adenosine A2B receptor and hyaluronan modulate pulmonary hypertension associated with chronic obstructive pulmonary disease. Am. J. Respir. Cell Mol. Biol. *49*, 1038–1047.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kölling, M., Haddad, G., Wegmann, U., Kistler, A., Bosakova, A., Seeger, H., Hübel, K., Haller, H., Mueller, T., Wüthrich, R.P., and Lorenzen, J.M. (2019). Circular RNAs in Urine of Kidney Transplant Patients with Acute T Cell-Mediated Allograft Rejection. Clin. Chem. *65*, 1287–1294.

Lefever, S., Pattyn, F., De Wilde, B., Coppieters, F., De Keulenaer, S., Hellemans, J., and Vandesompele, J. (2017). High-throughput PCR assay design for targeted resequencing using primerXL. BMC Bioinformatics *18*, 400.

Li, S., Li, Y., Chen, B., Zhao, J., Yu, S., Tang, Y., Zheng, Q., Li, Y., Wang, P., He, X., and Huang, S. (2018a). exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. Nucleic Acids Res. *46* (*D1*), D106–D112.

Li, X., Yang, L., and Chen, L.-L. (2018b). The Biogenesis, Functions, and Challenges of Circular RNAs. Mol. Cell *71*, 428–442.

Lin, H., Zhou, Q., Wu, W., and Ma, Y. (2019). Midkine Is a Potential Urinary Biomarker for Non-Invasive Detection of Bladder Cancer with Microscopic Hematuria. OncoTargets Ther. *12*, 11765–11775.

Liu, F., Ma, R., Wang, Y., and Zhang, L. (2018). The Clinical Importance of *Campylobacter concisus* and Other Human Hosted *Campylobacter* Species. Front. Cell. Infect. Microbiol. *8*, 243.

Liu, B., Song, F., Yang, Q., Zhou, Y., Shao, C., Shen, Y., Zhao, Z., Tang, Q., Hou, Y., and Xie, J. (2019). Characterization of tissue-specific biomarkers with the expression of circRNAs in forensically relevant body fluids. Int. J. Legal Med. *133*, 1321–1331.

Locati, M.D., Terpstra, I., de Leeuw, W.C., Kuzak, M., Rauwerda, H., Ensink, W.A., van Leeuwen, S., Nehrdich, U., Spaink, H.P., Jonker, M.J., et al. (2015). Improving small RNA-seq by using a synthetic spike-in set for size-range quality control together with a set for data normalization. Nucleic Acids Res. *43*, e89.

Lorenzi, L., Chiu, H.-S., Avila Cobos, F., Gross, S., Volders, P.-J., Cannoodt, R., Nuytens, J., Vanderheyden, K., Anckaert, J., Lefever, S., et al. (2019). The RNA Atlas, a single nucleotide resolution map of the human transcriptome. BioRxiv. https://doi.org/10.1101/807529.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Maker, A.V., Hu, V., Kadkol, S.S., Hong, L., Brugge, W., Winter, J., Yeo, C.J., Hackert, T., Büchler, M., Lawlor, R.T., et al. (2019). Cyst Fluid Biosignature to Predict Intraductal Papillary Mucinous Neoplasms of the Pancreas with High Malignant Potential. J. Am. Coll. Surg. *228*, 721–729.

Marzioni, M., Germani, U., Agostinelli, L., Bedogni, G., Saccomanno, S., Marini, F., Bellentani, S., Barbera, C., De Minicis, S., Rychlicki, C., et al. (2015). PDX-1 mRNA expression in endoscopic ultrasound-guided fine needle cytoaspirate: perspectives in the diagnosis of pancreatic cancer. Dig. Liver Dis. *47*, 138–143.

Max, K.E.A., Bertram, K., Akat, K.M., Bogardus, K.A., Li, J., Morozov, P., Ben-Dov, I.Z., Li, X., Weiss, Z.R., Azizian, A., et al. (2018). Human plasma and serum extracellular small RNA reference profiles and their clinical utility. Proc. Natl. Acad. Sci. USA *115*, E5334–E5343.

Memczak, S., Papavasileiou, P., Peters, O., and Rajewsky, N. (2015). Identification and Characterization of Circular RNAs As a New Class of Putative Biomarkers in Human Blood. PLoS ONE *10*, e0141214.

Metzenmacher, M., Váraljai, R., Hegedüs, B., Cima, I., Forster, J., Schramm, A., Scheffler, B., Horn, P.A., Klein, C.A., Szarvas, T., et al. (2020). Plasma Next Generation Sequencing and Droplet Digital-qPCR-Based Quantification of Circulating Cell-Free RNA for Noninvasive Early Detection of Cancer. Cancers (Basel) *12*, 353.

Mullen, K.M., and van Stokkum, I.H.M. (2012). nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). R package version 1.4. https://cran.r-project.org/web/packages/nnls/index.html.

Murakami, T., Yamamoto, C.M., Akino, T., Tanaka, H., Fukuzawa, N., Suzuki, H., Osawa, T., Tsuji, T., Seki, T., and Harada, H. (2018). Bladder cancer detection by urinary extracellular vesicle mRNA analysis. Oncotarget *9*, 32810–32821.

Murillo, O.D., Thistlethwaite, W., Rozowsky, J., Subramanian, S.L., Lucero, R., Shah, N., Jackson, A.R., Srinivasan, S., Chung, A., Laurent, C.D., et al. (2019). exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids. Cell *177*, 463–477.e15.

Oreo, K.M., Gibson, P.G., Simpson, J.L., Wood, L.G., McDonald, V.M., and Baines, K.J. (2014). Sputum ADAM8 expression is increased in severe asthma and COPD. Clin. Exp. Allergy *44*, 342–352.

Ounit, R., and Lonardi, S. (2016). Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics *32*, 3823–3825.

Pieragostino, D., Agnifili, L., Cicalini, I., Calienno, R., Zucchelli, M., Mastropasqua, L., Sacchetta, P., Del Boccio, P., and Rossi, C. (2017). Tear Film Steroid Profiling in Dry Eye Disease by Liquid Chromatography Tandem Mass Spectrometry. Int. J. Mol. Sci. *18*, 1349.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rozowsky, J., Kitchen, R.R., Park, J.J., Galeev, T.R., Diao, J., Warrell, J., Thistlethwaite, W., Subramanian, S.L., Milosavljevic, A., and Gerstein, M. (2019). exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling. Cell Syst. *8*, 352–357.e3.

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., and Walker, A.W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. *12*, 87.

Saugstad, J.A., Lusardi, T.A., Van Keuren-Jensen, K.R., Phillips, J.I., Lind, B., Harrington, C.A., McFarland, T.J., Courtright, A.L., Reiman, R.A., Yeri, A.S., et al. (2017). Analysis of extracellular RNA in cerebrospinal fluid. J. Extracell. Vesicles *6*, 1317577.

Simon, L.M., Karg, S., Westermann, A.J., Engel, M., Elbehery, A.H.A., Hense, B., Heinig, M., Deng, L., and Theis, F.J. (2018). MetaMap: an atlas of metatranscriptomic reads in human disease-related RNA-seq data. Gigascience *7*, giy070.

Singh Patidar, B., Meena, A., Kumar, M., Menon, B., Rohil, V., and Kumar Bansal, S. (2018). Adenosine Metabolism in COPD: A Study on Adenosine Levels, 5′-Nucleotidase, Adenosine Deaminase and Its Isoenzymes Activity in Serum, Lymphocytes and Erythrocytes. COPD *15*, 559–571.

Srinivasan, S., Yeri, A., Cheah, P.S., Chung, A., Danielson, K., De Hoff, P., Filant, J., Laurent, C.D., Laurent, L.D., Magee, R., et al. (2019). Small RNA Sequencing across Diverse Biofluids Identifies Optimal Methods for exRNA Isolation. Cell *177*, 446–462.e16.

Stankovic, M., Nikolic, A., Nagorni-Obradovic, L., Petrovic-Stanojevic, N., and Radojkovic, D. (2017). Gene-Gene Interactions Between Glutathione S-Transferase M1 and Matrix Metalloproteinases 1, 9, and 12 in Chronic Obstructive Pulmonary Disease in Serbians. COPD *14*, 581–589.

Strong, M.J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., Fewell, C., Taylor, C.M., and Flemington, E.K. (2014). Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. PLoS Pathog. *10*, e1004437.

Tian, Y., Li, L., Zhang, F., and Xu, J. (2016). Seminal plasma HSPA2 mRNA content is associated with semen quality. J. Assist. Reprod. Genet. *33*, 1079–1084.

Umu, S.U., Langseth, H., Bucher-Johannessen, C., Fromm, B., Keller, A., Meese, E., Lauritzen, M., Leithaug, M., Lyle, R., and Rounge, T.B. (2018). A comprehensive profile of circulating RNAs in human serum. RNA Biol. *15*, 242–250.

Vo, J.N., Cieslik, M., Zhang, Y., Shukla, S., Xiao, L., Zhang, Y., Wu, Y.-M., Dhanasekaran, S.M., Engelke, C.G., Cao, X., et al. (2019). The Landscape of Circular RNA in Cancer. Cell *176*, 869–881.e13.

Waller, R., Wyles, M., Heath, P.R., Kazoka, M., Wollff, H., Shaw, P.J., and Kirby, J. (2018). Small RNA Sequencing of Sporadic Amyotrophic Lateral Sclerosis Cerebrospinal Fluid Reveals Differentially Expressed miRNAs Related to Neural and Glial Activity. Front. Neurosci. *11*, 731.

Weber, J.S. (2017). Biomarkers for Checkpoint Inhibition. Am. Soc. Clin. Oncol. Educ. Book *37*, 205–209.

Weber, J.A., Baxter, D.H., Zhang, S., Huang, D.Y., Huang, K.H., Lee, M.J., Galas, D.J., and Wang, K. (2010). The microRNA spectrum in 12 body fluids. Clin. Chem. *56*, 1733–1741.

Weiland, M., Gao, X.-H., Zhou, L., and Mi, Q.-S. (2012). Small RNAs have a large impact: circulating microRNAs as biomarkers for human diseases. RNA Biol. *9*, 850–859.

Welch, R.A., Shaw, M.K., and Welch, K.C. (2016). Amniotic fluid LPCAT1 mRNA correlates with the lamellar body count. J. Perinat. Med. *44*, 531–532.

Yeri, A., Courtright, A., Reiman, R., Carlson, E., Beecroft, T., Janss, A., Siniard, A., Richholt, R., Balak, C., Rozowsky, J., et al. (2017). Total Extracellular Small RNA Profiles from Plasma, Saliva, and Urine of Healthy Subjects. Sci. Rep. *7*, 44061.

Yuan, T., Huang, X., Woodcock, M., Du, M., Dittmar, R., Wang, Y., Tsai, S., Kohli, M., Boardman, L., Patel, T., and Wang, L. (2016). Plasma extracellular RNA profiles in healthy and cancer patients. Sci. Rep. *6*, 19413.

Zhang, L., Farrell, J.J., Zhou, H., Elashoff, D., Akin, D., Park, N.-H., Chia, D., and Wong, D.T. (2010). Salivary transcriptomic biomarkers for detection of resectable pancreatic cancer. Gastroenterology *138*, 949–957.e7.

Zhang, X.-O., Dong, R., Zhang, Y., Zhang, J.-L., Luo, Z., Zhang, J., Chen, L.-L., and Yang, L. (2016). Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res. *26*, 1277–1287.

Zhou, Z., Wu, Q., Yan, Z., Zheng, H., Chen, C.-J., Liu, Y., Qi, Z., Calandrelli, R., Chen, Z., Chien, S., et al. (2019). Extracellular RNA in a single droplet of human serum reflects physiologic and disease states. Proc. Natl. Acad. Sci. USA *116*, 19200–19208.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
miRNeasy Serum/Plasma Kit	QIAGEN	Cat# 217184
External RNA Control Consortium (ERCC) spike-in controls	ThermoFisher Scientific	Cat# 4456740
TruSeq RNA Exome Library Prep Kit	Illumina	Cat# 20020189; Cat# 20020490; Cat# 20020183
TruSeq Small RNA Library Prep Kit	Illumina	Cat# RS-200-0012; Cat# RS-200-0024; Cat# RS-200-0036; Cat# RS-200-0048
KAPA Library Quantification Kit	Roche Diagnostics	Cat# KK4854
iScript Advanced cDNA Synthesis Kit	BioRad	Cat# 172-5038
Sso Advanced PreAmp Supermix	BioRad	Cat# 172-5160
SsoAdvanced Universal SYBR Green Supermix	BioRad	Cat# 172-5275
Deposited Data		
RNA sequencing data of discovery cohort and validation cohorts	This paper	EGAS00001003917
Processed RNA sequencing data of discovery cohort and validation cohorts	This paper	http://r2platform.com/HumanBiofluidRNAAtlas
mRNA capture sequencing of 36 cancerous tissue types obtained from the MiOncoCirc Database	University of Michigan Clinical Sequencing Exploratory Research	dbGaP phs000673.v3.p1
Software and Algorithms		
FastQC software	Babraham Institute	RRID:SCR_014583
STAR (v2.6.0) Ultrafast Universal RNA-seq Aligner	Dobin et al., 2013	RRID:SCR_015899
HTSeq (v0.9.1)	EMBL Heidelberg	RRID:SCR_005514
Cutadapt (v1.8.1)	National Bioinformatics Infrastructure Sweden	RRID:SCR_011841
Bowtie (v1.1.2)	Johns Hopkins University	RRID:SCR_005476
R package DESeq2 (v1.20.0)	Love et al., 2014	http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html
qbase+ software	Biogazelle	www.qbaseplus.com
MetaMap pipeline	Simon et al., 2018	https://github.com/theislab/MetaMap
exceRpt small RNA-seq pipeline	Genboree Workbench	http://genboree.org/index
Circexplorer2	Zhang et al., 2016	https://circexplorer2.readthedocs.io/en/latest/
R package ggplot2	https://ggplot2.tidyverse.org/	https://cran.r-project.org/web/packages/ggplot2/index.html
Other		
Sequin spike-in controls	Garvan Institute of Medical Research	https://www.sequinstandards.com
RNA extraction Control (RC) spike-in controls	Integrated DNA Technologies	N/A
Library Prep Control (LP) spike-in controls	Integrated DNA Technologies	N/A

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Pieter Mestdagh (Pieter.Mestdagh@ugent.be).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The raw RNA-sequencing data have been deposited at the European Genome-phenome Archive (EGA):EGAS00001003917. The small RNA sequencing data was deposited in the exRNA Atlas portal (https://exrna-atlas.org). All spike-normalized sequencing data can be readily explored in the interactive web-based application R2: Genomics analysis and visualization platform (http://r2.amc.nl), and via a dedicated accessible portal (http://r2platform.com/HumanBiofluidRNAAtlas). This portal allows the analysis and visualization of mRNA, circRNA and miRNA abundance, as illustrated in Figure S13. All samples can be used for correlation, principle component, and gene set enrichment analyses, and many more. All other data are available within the article and Supplemental Information. The R scripts to reproduce the analyses and plots reported in this paper are available from the corresponding authors upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Donor material, collection and biofluid preparation procedure

Sample collection for the discovery cohort and sputum collection for the case/control cohort was approved by the ethics committee of Ghent University Hospital, Ghent, Belgium (no. B670201734450) and written informed consent was obtained from all donors according to the Helsinki declaration. Breast milk, colostrum, plasma, serum, sputum, seminal plasma, sweat, stool, tears and urine were obtained in healthy volunteers. All other biofluids were collected from non-oncological patients.

The collection of two case series of each 12 cases and 12 control samples was approved by the Masaryk Memorial Cancer Institute, Brno, Czech Republic (no. 14-08-27-01 and no. MOU190814). Urine was collected in healthy donors and muscle-invasive bladder cancer patients; CSF was collected in hydrocephalus patients and glioblastoma patients.

Collection of saliva samples in 12 healthy donors and in patients with diabetes mellitus for the case/control cohort was approved by the ethics committee of the Medical University of Vienna, Vienna, Austria (no. 2197/2015). Written informed consent was obtained from all donors. The demographic and clinical patient information is provided in Table S1. Detailed information on the sample collection per biofluid is provided in Methods S1. All samples, except tear fluid, plasma and serum, were centrifuged at 2000 g (*rcf.*) for 10 minutes without brake at room temperature. All samples were processed within 2 hours after collection. The cell-free supernatant was carefully pipetted into 2 mL LoBind tubes (Eppendorf LoBind microcentrifuge tubes, Z666556-250EA) and stored at −80°C.

METHOD DETAILS

RNA isolation and gDNA removal

RNA isolation from all biofluids, except tears

In the discovery cohort, two RNA isolations per biofluid and per sample were simultaneously performed by two researchers (E.V.E. and E.H.). In the end, RNA obtained from both RNA isolations was pooled per biofluid and per sample and this pooled RNA was used as starting material for both library preparations. Hence, small RNA and mRNA capture sequencing on the discovery cohort were performed on the same batch of RNA. In the case/control cohorts, one RNA isolation was performed per sample and the RNA was used as starting material for mRNA capture sequencing.

RNA was isolated with the miRNeasy Serum/Plasma Kit (QIAGEN, Hilden, Germany, 217184) according to the manufacturer's instructions. An input volume of 200 μL was used for all samples, except for tear fluid, and total RNA was eluted in 12 μL of RNase-free water. Tear fluid was collected with Schirmer strips and RNA was isolated directly from the strips (see further). Per 200 μL biofluid input volume, 2 μL Sequin spike-in controls (Garvan Institute of Medical Research) and 2 μl RNA extraction Control (RC) spike-ins (Integrated DNA Technologies) (Locati et al., 2015) were added to the lysate for TruSeq RNA Exome Library Prep sequencing and TruSeq Small RNA Library Prep sequencing, respectively. Details on the spike-in controls are available in the Methods S1.

Briefly, 2 μl External RNA Control Consortium (ERCC) spike-in controls (ThermoFisher Scientific, Waltham, MA, USA, 4456740), 2 μl Library Prep Control (LP) spike-ins (Integrated DNA Technologies) (Hafner et al., 2011), 1 μl HL-dsDNase and 1.6 μl reaction buffer were added to 12 μl RNA eluate, and incubated for 10 min at 37°C, followed by 5 min at 55°C. Per biofluid and per donor the RNA after gDNA removal was pooled. RNA was stored at −80°C and only thawed on ice immediately before the start of the library prep. Multiple freeze/thaw cycles did not occur.

RNA isolation from tear fluid

Tear fluid was collected in 8 healthy donors with Schirmer strips (2 strips per eye per donor), as previously described (Green-Church et al., 2008; Pieragostino et al., 2017). RNA was isolated within two hours after tear collection with the miRNeasy Serum/Plasma Kit (QIAGEN, Hilden, Germany, 217184), starting from one 2 mL tube containing each 4 Schirmer strips. The same reagent volumes as suggested by the manufacturer for a 200 μL input volume were used. Throughout the RNA isolation protocol, the two final RNA samples each result from 4 tear fluid samples (each containing the 4 strips of a single donor) that were pooled in a two-step method. First, the upper aqueous phase of two tear fluid samples was put together (in step 8 of the RNA isolation protocol). Second, the RNA eluate of these two samples was pooled into the final RNA that was used as input for the library prep (in step 15 of the RNA isolation protocol).

TruSeq RNA Exome library prep sequencing

Messenger RNA capture based libraries were prepared starting from 8.5 μL DNase treated and spike-in supplemented RNA eluate using the TruSeq RNA Exome Library Prep Kit (Illumina, San Diego, CA, USA). Each sample underwent individual enrichment according to the manufacturer's protocol. The quality and yield of the prepared libraries were assessed using a high sensitivity Small DNA Fragment Analysis Kit (Agilent Technologies, Santa Clara, CA, USA) according to manufacturer's instructions. The libraries were quantified using qPCR with the KAPA Library Quantification Kit (Roche Diagnostics, Diegem, Belgium, KK4854) according to manufacturer's instructions. Based on the qPCR results, equimolar library pools were prepared.

Paired-end sequencing was performed on a NextSeq 500 instrument using a high output v2 kit (Illumina, San Diego, CA, USA) with a read length of 75 nucleotides to an average sequencing depth of 11 million read pairs in the discovery cohort, 16.8 million read pairs in the sputum case/control cohorts, 15.4 million read pairs in the urine case/control cohort, 15 million read pairs in the CSF case/control cohort and 18.8 million read pairs in the saliva case/control cohort. Samples from the discovery cohort were randomly assigned over two pools and sequenced with a loading concentration of 1.2 pM (5% PhiX) and 1.6 pM (5% PhiX), respectively. Urine, CSF and saliva samples from the case/control cohorts were loaded in 3 separate runs at 2 pM (2% PhiX) and sputum samples from the case/control cohorts were loaded at 1.6 pM (5% PhiX).

TruSeq Small RNA library prep sequencing

Small RNA libraries were prepared starting from 5 μL DNase treated and spike-in supplemented RNA eluate using a TruSeq Small RNA Library Prep Kit (Illumina, San Diego, CA, USA) according to the manufacturer's protocol with two minor modifications(1). The RNA 3′ adaptor (RA3) and the RNA 5′ adaptor (RA5) were 4-fold diluted with RNase-free water(2) and the number of PCR cycles was increased to 16.

First, a volume-based pool of all 46 samples of the discovery cohort was sequenced. After PCR amplification, quality of libraries was assessed using a high sensitivity Small DNA Fragment Analysis Kit (Agilent Technologies, Santa Clara, CA, USA) according to manufacturer's instructions. Size selection of the pooled samples was performed using 3% agarose dye-free marker H cassettes on a Pippin Prep (Sage Science, Beverly, MA, USA) following manufacturer's instructions with a specified collection size range of 125–163 bp. Libraries were further purified and concentrated by ethanol precipitation, resuspended in 10 μl of 10 mM tris-HCl (pH = 8.5) and quantified using qPCR with the KAPA Library Quantification Kit (Roche Diagnostics, Diegem, Belgium, KK4854) according to manufacturer's instructions. The pooled library was quality controlled via sequencing at a concentration of 1.7 pM with 35% PhiX on a NextSeq 500 using a mid-output v2 kit (single-end 75 nucleotides, Illumina, San Diego, CA, USA), resulting in an average sequencing depth of 1 million reads, ranging from 3341 reads to 14 million reads. Twenty-three samples with less than 200 000 reads were assigned to a low concentrated pool, 23 samples with more than 17 million reads were assigned to a highly concentrated pool. Based on the read numbers from the mid output run, two new equimolar pools were prepared, purified and quantified as described higher. Both re-pooled libraries were then sequenced at a final concentration of 1.7 pM with 25% PhiX on a NextSeq 500 using a high output v2 kit (single-end, 75 nucleotides, Illumina, San Diego, CA, USA), resulting in an average sequencing depth of 9 million reads (range 817 469 – 41.7 million reads).

RT-qPCR

To validate findings observed in the RNA sequencing data, we performed a targeted mRNA and circRNA expression profiling with RT-qPCR for 8 differentially expressed mRNAs in sputum (COPD versus healthy control) and for the 10 most abundant circRNAs in sputum. As reference RNAs for normalization purposes, we selected Sequin spikes stably detected in all samples based on the available RNA sequencing data. The assays to measure mRNA, circRNA and Sequin spike expression were custom designed using primerXL (Lefever et al., 2017) (Table S1) and purchased from Integrated DNA Technologies, Inc. (Coralville, USA).

For cDNA synthesis, 5 μL of total RNA was reverse transcribed using the iScript Advanced cDNA Synthesis Kit (BioRad, USA, 172-5038) in a 10 μL volume. 5 μL of cDNA was pre-amplified in a 12-cycle PCR reaction using the SsoAdvanced PreAmp Supermix (Bio-Rad, USA, 172-5160) in a 50 μL reaction. Pre-amplified cDNA was diluted (1:8) and 2 μL was used as input for a 45-cycle qPCR reaction, quantifying 8 mRNAs and 10 circRNAs of interest with the SsoAdvanced Universal SYBR Green Supermix (BioRad, USA, 172-5275). All reactions were performed in 384-well plates on the LightCycler480 instrument (Roche) in a 5 μL reaction volume using 250 nM primer concentrations. Cq-values were determined with the LightCycler®480 Software (release 1.5.0, Roche) with the "Abs Quant/2nd Derivative Max" method.

The geNorm analysis to select the optimal number of reference targets was performed using Biogazelle's qbase+ software (www.qbaseplus.com) using log2-transformed RNA count data. We observed medium reference target stability (average geNorm M $\leq$ 1.0) with an optimal number of reference targets in this experimental situation of two (geNorm V $<$ 0.15 when comparing a normalization factor based on the two or three most stable targets). As such, the optimal normalization factor can be calculated as the geometric mean of reference targets R2_150 and R2_65. These Sequin spike RNAs were considered as reference RNAs.

QUANTIFICATION AND STATISTICAL ANALYSIS

Processing TruSeq RNA Exome sequencing data

Read quality was assessed by running FastQC (v0.11.5) on the FASTQ files and reads shorter than 35 nucleotides and with a quality (phred) score $<$ 30 were removed. The reads were mapped with STAR (v2.6.0). Mapped reads were annotated by matching genomic coordinates of each read with genomic locations of mRNAs (obtained from UCSC GRCh38/hg38 and Ensembl, v91) or by matching the spike-in sequences. Picard (v2.18.5) was used for duplicate removal. HTSeq (v0.9.1) was used for quantification of PCR deduplicated reads. A cut-off for filtering noisy genes was set based on historic data to remove noisy genes. Using a threshold of 4 counts, at least 95% of the single positive replicate values are filtered out. A table with the read count of mRNAs per sample is provided in Table S1.

Processing TruSeq Small RNA sequencing data

Adaptor trimming was performed using Cutadapt (v1.8.1) with a maximum error rate of 0.15. Reads shorter than 15 nts and those in which no adaptor was found were discarded. For quality control the FASTX-Toolkit (v0.0.14) was used, a minimum quality score of 20 in at least 80% of nucleotides was applied as a cutoff. The reads were mapped with Bowtie (v1.1.2) without allowing mismatches. Mapped reads were annotated by matching genomic coordinates of each read with genomic locations of miRNAs (obtained from miRBase, v22) and other small RNAs (obtained from UCSC GRCh38/hg38 and Ensembl, v91) or by matching the spike-in sequences. Reads that map to more than one small RNA biotype are considered as ‘‘multimapped.’’ Only reads that uniquely map to one biotype, are included in the different small RNA biotype categories in Figure 2. Reads assigned as ‘‘not annotated’’ represent mapped reads that could not be attributed to small RNA biotype groups. The mapping locations of the uniquely mapped not annotated reads were cross-matched with all exonic, intronic and intergenic positions in the Ensembl reference transcriptome (obtained from UCSC GRCh38/hg38 and Ensembl, v91) using the intersect feature in bedtools (requiring an overlap of 100% between the not annotated read and the position of the Ensembl reference transcriptome). As for the mRNA data, genes with fewer than 4 counts were filtered out. A table with the read count of miRNAs per sample is provided in Table S1.

Calculation of endogenous RNA concentration

A biofluid volume-based normalization was applied by dividing the number of RNA reads consumed by the endogenous transcripts by the sum of the Sequin reads for mRNA data and by the sum of the RC reads for small RNA data. The spike-normalized data represent relative abundance values of RNA molecules proportional to the input volume.

The mass of endogenous mRNA present in 1 mL of each biofluid was estimated based on the read count for ERCC-00130 spike-in RNA, detected in all biofluids with at least 4 counts. The RNA eluate of each sample contains 6E-20 mol of ERCC-00130, which has a molecular weight of 340415.55 g/mol. Based on the read count for the endogenous mRNA, the corresponding mass of endogenous mRNA in the eluate was calculated and corrected for input volume.

Exogenous RNA characterization

The exogenous RNA content in the mRNA data was assessed using the MetaMap pipeline (Simon et al., 2018). Briefly, all reads were mapped to the human reference genome (hg38) using STAR (v2.5.2) (Dobin et al., 2013). Unmapped reads were subsequently subjected to metagenomic classification using CLARK-S (v1.2.3) (Ounit and Lonardi, 2016). Reads were summed across all bacterial species.

The exogenous RNA content in the small RNA data was assessed using the exceRpt small RNA-seq pipeline (v4.6.2) in the Genboree workbench with default settings (Rozowsky et al., 2019). Briefly, after adaptor trimming, read quality was assessed by FASTQC (v0.11.2). A minimum quality score of 20 in at least 80% of nucleotides was applied as cutoff. The minimum read length after adaptor trimming was set to 18 nucleotides. Reads were first mapped to the custom spike-in sequences using Bowtie2 (v2.2.6), followed by mapping the unmapped reads with STAR (v2.4.2a) to UniVec contaminants and human ribosomal (rRNA) sequences to exclude them before mapping (also with STAR) to the following databases: miRbase (v21), gtRNAdb, piRNABank, GeneCode version 24 (hg38) and circBase (version last updated in July 2017). A single mismatch was allowed during mapping to the human genome. Unmapped reads were then mapped with STAR to exogenous miRNAs and rRNAs. In the end, the remaining unmapped reads were mapped to the genomes of all sequenced species in Ensembl and NCBI. No mismatches were allowed during exogenous alignment. Raw read counts obtained from the Genboree workbench were further analyzed in R (v3.5.1) making use of tidyverse (v1.2.1).

Circular RNA detection and circular/linear ratio determination

Only TruSeq RNA Exome reads passing quality control (base calling accuracy of $\geq$ 99% in at least 80% of the nucleotides in both mates of a pair) were included in this analysis. Clumpify dedupe (v38.26) was used to remove duplicates in paired-end mode (2 allowed substitutions, kmer size of 31 and 20 passes). We used a two-step mapping strategy to identify forward splice (further referred to as linear) junction reads and backsplice junction reads. First, reads were aligned with TopHat2 (v2.1.0) to the GRCh38/hg38 reference genome (Ensembl, v91) (Kim et al., 2013). Micro-exons were included, a minimum anchor length of 6 nucleotides was required, and up to two mismatches in the anchor region were allowed. The resulting output contains linear junction information. Second, unmapped reads from the first mapping strategy were realigned with TopHat2 (v2.1.0) to the same reference, but this time with the fusion search option that can align reads to potential fusion transcripts. Processing the fusion search output with CIRCexplorer2 parse (v2.3.3) results in backsplice junction information (Zhang et al., 2016). Junction read counts obtained with the mapping strategies described above were used as a measure for the relative level of linear and circular RNA in each sample. Only genes with at least one detected backsplice junction were considered. Junctions that could be part of both linear and circular transcripts (ambiguous junctions) were filtered out. As there is currently no consensus on how to calculate the circular to linear ratio (CIRC/LIN), we decided to calculate the ratio in two different ways (Figure S7). The circRNA fraction is defined as 100*CIRC/(CIRC+LIN). The first method (referred to as ‘‘backsplice junction-level method’’) zooms in on each particular backsplice junction. CIRC was defined as the backsplice junction read count of one particular backsplice junction. LIN was defined as the average read count of all junctions flanking the backsplice junction of interest. The second method (referred to as ‘‘gene-level method’’) considers all backsplice junctions within a given gene. CIRC was defined as the average number of backsplice junction reads for a given gene. LIN was defined as the average number of linear junction reads for a given gene. For both methods, CIRC > 3 was used as a cut-off for filtering noisy backsplice junctions. To enable a comparison of the circular/linear genic ratios in biofluids with those of tissues, the mRNA capture sequencing FASTQ files of 16 cancerous tissue types (34 samples in total) were downloaded from the MiOncoCirc database (dbGaP Study Accession phs000673.v3.p1) (Vo et al., 2019). A list with the downloaded samples is attached in Table S1. A table with the read count of backsplice junctions per sample is provided in Table S1.

Assessment of tissue and cell contribution to biofluid exRNA

Using total RNA-sequencing data from 27 normal human tissue types and 5 immune cell types from peripheral blood from the RNA Atlas (Lorenzi et al., 2019), we created gene sets containing marker genes for each individual entity (Table S1). We removed redundant tissues and cell types from the original RNA Atlas (e.g., granulocytes and monocytes were present twice; brain was kept and specific brain sub-regions such as cerebellum, frontal cortex, occipital cortex and parietal cortex were removed) and we used genes where at least one tissue or cell type had expression values greater or equal to 1 TPM normalized counts. A gene was considered to be a marker if its abundance was at least 5 times higher in the most abundant sample compared to the others. For the final analysis, only tissues and cell types with at least 3 markers were included, resulting in 26 tissues and 5 immune cell types.

Gene abundance read counts from the biofluids were normalized using Sequin spikes as size factors in DESeq2 (v1.22.2). For all marker genes within each gene set, we computed the log2 fold changes between the median read count of a biofluid sample pair versus the median read count of all other biofluids. The median log2 fold change of all markers in a gene set was selected, followed by z-score transformation over all biofluids (Figure 6). For visualization purposes, only tissues and cell types with a z-score $\geq |1|$ in at least one biofluid were used.

Cellular deconvolution of pancreatic cyst fluid samples

To build the reference matrix for the computational deconvolution of pancreatic cyst fluid samples, single cell RNA sequencing data of 10 pancreatic cell types (Baron et al., 2016) was processed with the statistical programming language R (v3.6.0). For each gene, the mean count across all individual cells from each cell type was computed. Next, this reference matrix was normalized using the trimmed means of M values (TMM) with the edgeR package (v3.26.4)(Robinson et al., 2010). Limma-voom (v3.40.2) (Ritchie et al., 2015) was used for subsequent differential gene expression analysis and those genes with an absolute fold change greater or equal to 2 and an adjusted p value < 0.05 (Benjamini-Hochberg) were retained as markers (Benjamini and Hochberg, 1995). Finally, using these markers and both the pancreatic cyst fluid samples and the reference matrix described above, the cell type proportions were obtained through computational deconvolution using non-negative least-squares (nnls package; v1.4) (Avila Cobos et al., 2018; Mullen and van Stokkum, 2012).

Differential expression analysis in case/control cohorts

Further processing of the count tables was done with R (v3.5.1) making use of tidyverse (v1.2.1). Gene abundance expression read counts from the biofluids were normalized using the sum of all reads mapping to Sequin spikes as size factors in DESeq2 (v1.20.0) (Love et al., 2014). To assess the biological signal in the case/control cohorts, we performed differential expression analysis between the patients and control groups using DESeq2 (v1.20.0). Genes were considered differentially expressed when the absolute log2 fold change > 1 and at q < 0.05. A list with differentially expressed genes in all case/control cohorts can be found in Table S1.